IDA PAPER P-2240

# CLASSIFICATION UTILITY: MEASURING AND IMPROVING BENEFITS IN MATCHING PERSONNEL TO JOBS

Cecil D. Johnson
Joseph Zeidner



April 1990

*Prepared for*
Office of the Under Secretary of Defense for Acquisition
(Research and Advanced Technology)



INSTITUTE FOR DEFENSE ANALYSES
1801 N. Beauregard Street, Alexandria, Virginia 22311-1772

IDA Log No. HQ 89-34688

## DEFINITIONS

IDA publishes the following documents to report the results of its work.

### Reports

Reports are the most authoritative and most carefully considered products IDA publishes. They normally embody results of major projects which (a) have a direct bearing on decisions affecting major programs, (b) address issues of significant concern to the Executive Branch, the Congress and/or the public, or (c) address issues that have significant economic implications. IDA Reports are reviewed by outside panels of experts to ensure their high quality and relevance to the problems studied, and they are released by the President of IDA.

### Group Reports

Group Reports record the findings and results of IDA established working groups and panels composed of senior individuals addressing major issues which otherwise would be the subject of an IDA Report. IDA Group Reports are reviewed by the senior individuals responsible for the project and others as selected by IDA to ensure their high quality and relevance to the problems studied, and are released by the President of IDA.

### Papers

Papers, also authoritative and carefully considered products of IDA, address studies that are narrower in scope than those covered in Reports. IDA Papers are reviewed to ensure that they meet the high standards expected of refereed papers in professional journals or formal Agency *reports*.

### Documents

IDA Documents are used for the convenience of the sponsors or the analysts (a) to record substantive work done in quick reaction studies, (b) to record the proceedings of conferences and meetings, (c) to make available preliminary and tentative results of analyses, (d) to record data developed in the course of an investigation, or (e) to forward information that is essentially unanalyzed and unevaluated. The review of IDA Documents is suited to their content and intended use.

The work reported in this document was conducted under contract MDA 903 89 C 0003 for the Department of Defense. The publication of this IDA document does not indicate endorsement by the Department of Defense, nor should the contents be construed as reflecting the official position of that Agency.

This Paper has been reviewed by IDA to assure that it meets high standards of thoroughness, objectivity, and appropriate analytical methodology and that the results, conclusions and recommendations are properly supported by the material presented.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



**1. AGENCY USE ONLY** *(Leave blank)*

**2. REPORT DATE**
April 1990

**3. REPORT TYPE AND DATES COVERED**
Final--September 1988 to April 1990

**4. TITLE AND SUBTITLE**
Classification Utility: Measuring and Improving Benefits in Matching Personnel to Jobs

**5. FUNDING NUMBERS**
C - MDA 903 89 C 0003
T - T-D2-435

**6. AUTHOR(S)**
Cecil D. Johnson, Joseph Zeidner

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311-1772

**8. PERFORMING ORGANIZATION REPORT NUMBER**
IDA Paper P-2240

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
OUSD(A)/R&AT
The Pentagon, Room 3D129
Washington, DC 20301

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**
Approved for public release; distribution unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

Over a period of two decades, the content of both test composites and the operational test battery, the Armed Services Vocational Battery (ASVAB), have been selected to maximize predictive validity with little attention given to improving the classification efficiency of the total set of test composites in a multi-job, optimal assignment situation. This emphasis on predictive validity and its operational simplicity can be shown to be fundamentally erroneous with respect to both empirical results and psychometric theory.

Although the present composites are of marginal value, considerable classification efficiency is potentially obtainable from the existing ASVAB if it is used in accordance with differential assignment principles. The primary objective of this report is to describe the principles underlying selection and classification for multiple jobs identified through reliance on the measurement of mean predicted performance.

The report includes descriptions of: (1) a new taxonomy for the total "personnel utilization decision process" that embraces selection, classification and placement; (2) methods of measuring potential classification efficiency and its components; (3) techniques for improving classification efficiency through selecting predictors or structuring job families and associated full least squares composites; (4) a means of identifying rotations of principal component factors that maximize the differential efficiency of a battery; and (5) a practical and flexible model sampling and simulation approach as a tool for measuring selectional and classification utility.

**14. SUBJECT TERMS**
military personnel classification, aptitude testing, job performance, manpower, classification, classification utility, utility analysis

**15. NUMBER OF PAGES**
256

**16. PRICE CODE**

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| UNCLASSIFIED | UNCLASSIFIED | UNCLASSIFIED | SAR |

IDA PAPER P-2240

# CLASSIFICATION UTILITY: MEASURING AND IMPROVING BENEFITS IN MATCHING PERSONNEL TO JOBS

Cecil D. Johnson
Joseph Zeidner

April 1990

INSTITUTE FOR DEFENSE ANALYSES

Contract MDA 903 89 C 0003
Task T-D2-435

# ACKNOWLEDGMENTS

We wish to express our sincere appreciation for support of this study by making the Army 1984 recruit accession base available and by arranging for the simulations required for the utility analysis to:

Edgar Johnson, Technical Director, The Army Research Institute (ARI),

Newall Eaton, Director, Manpower and Personnel Research Laboratory, ARI,

Curtis Gilroy, Chief, Manpower and Personnel Policy Research Group, ARI.

The efforts of Arthur Drucker, formerly at the Army Research Institute, in editing the manuscript are gratefully acknowledged.

We are grateful for the reviews and useful technical comments made by:

Hubert Brogden, Professor Emeritus, Purdue University

Wayne Cascio, University of Colorado at Denver

Curtis Gilroy, The Army Research Institute

Paul Horst, Professor Emeritus, University of Washington

Edgar Johnson, The Army Research Institute

Robert Sadacca, HumRRO

William Sands, Navy Personnel Research and Development Center

Frank Schmidt, The University of Iowa

Richard Sorenson, Navy Personnel Research and Development Center.

This study was performed for the Office of the Under Secretary of Defense (Acquisition)/(Research and Advanced Technology). Technical cognizance for the work, under Task Order T-D2-435, was assigned to the Assistant for Training and Personnel Technology. We are indebted to Earl Alluisi for his contributions to and interest in the effort, and to Jesse Orlansky, Institute for Defense Analyses, for suggesting that the utility of selection and classification procedures be evaluated, for providing valuable information, and for his many useful suggestions on preparing this report.

# ABSTRACT


Over a period of two decades, the content of both test composites and the operational test battery, the Armed Services Vocational Battery (ASVAB), have been selected to maximize predictive validity with little attention given to improving the classification efficiency of the total set of test composites in a multi-job, optimal assignment situation. This emphasis on predictive validity and its operational simplicity can be shown to be fundamentally erroneous with respect to both empirical results and psychometric theory.

Although the present composites are of marginal value, considerable classification efficiency is potentially obtainable from the existing ASVAB if it is used in accordance with differential assignment principles. The primary objective of this report is to describe the principles underlying selection and classification for multiple jobs identified through reliance on the measurement of mean predicted performance.

The report includes descriptions of: (1) a new taxonomy for the total "personnel utilization decision process" that embraces selection, classification and placement; (2) methods of measuring potential classification efficiency and its components; (3) techniques for improving classification efficiency through selecting predictors or structuring job families and associated full least squares composites; (4) a means of identifying rotations of principal component factors that maximize the differential efficiency of a battery; and (5) a practical and flexible model sampling and simulation approach as a tool for measuring selectional and classification utility.

# CONTENTS

# FIGURES

# TABLES

# ABBREVIATIONS

| | |
|---|---|
| AA | Aptitude Area |
| ACB | Army Classification Battery |
| AFQT | Armed Forces Qualification Test |
| AGCT | Army General Classification Test |
| AI | Aptitude Index |
| AR | Arithmetic Reasoning |
| AS | Auto Shop Information |
| ASVAB | Armed Services Vocational Aptitude Battery |
| AVF | All-Volunteer Force |
| CL | Clerical/Administrative |
| CO | Combat |
| CREPID | Cascio-Ramos Estimate of Performance In Dollars |
| CS | Coding Speed |
| DEP | Delayed Entry Program |
| DI | Disposition Index |
| EI | Electronic Information |
| EL | Electronic Repair |
| EPAS | Enlisted Personnel Allocation System |
| F | Finger Dexterity |
| FA | Field Artillery |
| FLS | Full Least Squares |
| g | General Component |
| G | General Intelligence |
| GATB | General Aptitude Test Battery |
| GM | General Maintenance |

| | |
|---|---|
| GS | General Science |
| GVN | Cognitive Ability |
| $H_a$ | Absolute Validity Index |
| $H_d$ | Differential Validity Index |
| HRA | Human Resource Accounting |
| LSE | Least Squares Estimate |
| LP | Linear Program |
| Max-PSE | Maximum Potential Selection Efficiency |
| MC | Mechanical Comprehension |
| MDS | Multidimensional Screening |
| MK | Mathematical Knowledge |
| MM | Mechanical Maintenance |
| MOS | Military Occupational Speciality |
| MOSLS | Military Occupational Speciality Level System |
| MPP | Mean Predicted Performance |
| N | Numerical Aptitude |
| NO | Numerical Operations |
| OAE | Operational Allocation Efficiency |
| OF | Operators/Food |
| OPM | Office of Personnel Management |
| OPTAACL | Optimization on Aptitude Area Score, Classification Only |
| OPTAASC | Optimization on Aptitude Area Score, Selection and Classification |
| OPTFLS | Optimal Assignment, Full Least Squares Prediction |
| OPTPRFCL | Optimization on Single Composite Predicted Performance, Classification Only |
| P | Form Perception |
| PAE | Potential Allocation Efficiency |
| PACE | Professional and Administrative Career Examination |
| PAT | Programmer Aptitude Test |

| | |
|---|---|
| PC | Paragraph Comprehension |
| PCE | Potential Classification Efficiency |
| PDI | Point Distance Index |
| PSE | Potential Selection Efficiency |
| PUE | Personnel Utilization Efficiency |
| Q | Clerical Perception |
| RDO | Radial Drill Operator |
| S | Spatial Aptitude |
| SC | Surveillance/Communications |
| *SDy* | Standard Deviation of Performance in Dollar Terms |
| SIMPO | Simulation of Personnel Operations |
| SQT | Skill Qualification Test |
| SR | Selection Ratio |
| SRQ | Perceptual Ability |
| ST | Skilled Technical |
| TC | Tank Commander |
| u | Unique Component |
| USAREC | U.S. Army Recruiting Command |
| V | Verbal Aptitude |
| VE | General Verbal Ability |
| WK | Word Knowledge |

# SUMMARY

## A. PURPOSE

The current operational Army personnel classification and person-job matching system utilizes a set of nine aptitude area test composites corresponding to nine job families that evolved from two decades of research emphasis on enhancing predictive validity. The content of both test composites and the operational test battery, the Armed Services Vocational Aptitude Battery (ASVAB), was selected to maximize predictive validity with little or no attention paid to improving the classification efficiency of the total set of test composites in a multi-job, optimal assignment situation. Traditionally, the number of tests per composite has been kept small and the weights restricted to unity--or, at most, to two or three--in order to simplify the operational use of the composites. This emphasis on predictive validity and its operational simplicity (required in a precomputer age) can be shown to be either outdated or fundamentally erroneous with respect to both empirical results and psychometric theory.

Although the present ASVAB composites are of marginal value, considerable classification efficiency is potentially obtainable from the existing ASVAB if it is used in accordance with differential assignment principles. The primary objective of this report then is to describe the principles underlying selection and classification for multiple jobs identified through reliance on the measurement of mean predicted performance (MPP). The report embraces the total personnel utilization process and focuses on techniques for measuring and improving classification efficiency.

In a companion study (Zeidner and Johnson, 1989b), we estimated that implementing the tenets of differential assignment theory described in this report would bring about a large aggregate gain in MPP. Our "ball park" estimate of gains attributable to improved operational procedures to increase potential classification efficiency (PCE) exceeds 200 percent, or four-tenths of a standard deviation. We predicted that the largest contribution to PCE gains are full least squares (FLS) predictor composites; next are enlarged and restructured job families; and then the addition of classification efficient tests in the battery. We know from our simulation results that improvements of one- or two-tenths of a standard deviation could be worth well over 200 million dollars annually to the Army.

## B. MEASURING CLASSIFICATION EFFICIENCY

We begin with a taxonomy of personnel classification. The purpose of personnel classification is to match individuals and jobs in a manner that maximizes aggregate performance. Such classification decisions are a major concern in the military services and are of increasing interest in industry and in student counseling. We refer to the implementation of classification decisions as the "assignment process;" our generic term for the matching of individuals to either jobs (i.e., military occupational speciality) or to a level within a job (placement) is "assignment." In our taxonomy of personnel utilization processes "assignment" is subdivided into "classification" and "placement," and "classification" is further subdivided into "hierarchical classification" and "allocation."

Traditionally, in selection and placement, only a single job is involved, and can be accomplished with one or more predictors. The outcome is determined by an individual's position along a single predicted performance continuum. Classification decisions provide the basis for assigning a selected pool of individuals to more than one job. As in selection, these assignments can be made on the basis of a single predictor continuum adjusted to predict performance by reflecting job validities and/or values. When the predictors are adjusted in such a manner that the mean adjusted predictor scores and the mean criterion scores have the same rank order across jobs, a hierarchical layering effect that makes a positive contribution to the benefits obtainable from classification is evident. A hierarchical layering effect due to either a variation across jobs of the validities of job specific test composites, or to the value assigned to each job and reflected in predictor score means and/or variances, assures that the assignment process is, at least in part, influenced by hierarchical classification.

Classification that does not capitalize on hierarchical layering effects is referred to as "allocation." While hierarchical classification can be unidimensional (e.g., based entirely on a single predictor), allocation requires multiple predictors measuring more than one dimension in the joint predictor-criterion space. Validity is determined individually against each job's performance criterion; the set of job criteria should also be multidimensional. Thus a classification battery requires a separate assignment variable (criterion specific composite) for each criterion, if allocation efficiency is to be maximized. In practice, a smaller number of tests than are in the total battery are often used rather than in the LSEs (least squares estimates) from the total battery, the complete regression equation for all predictors. The particular combination of predictors employed out of the total battery plus the specifi weight given each predictor varies with each job criterion. In the Army, for example, a different unit-weighted, three-test combination or aptitude area composite currently is used in assigning individuals to jobs in each of nine families.

It is often assumed that the utility of the classification process is a direct function of differential validity. More precisely, differential validity is the level of prediction, using full battery LSEs, of differences among criterion scores. We also use the term in reference to the validity vector for a job having differential validity, i.e., being more valid for its own job family than for any other job family. Unfortunately, a simulation study is required to translate the effect of differential validity into mean predicted performance (MPP), that in turn can be readily translated into utility. The utility of a classification battery can be characterized as being directly proportional to the average predicted performance of incumbents in a number of different jobs after optimal assignment process has been used with quotas taken into account.

When the test content of the selection/classification battery has been fully determined and only the selection of test composites and weights for use in the selection and/or classification of applicants for each job remains to be determined, the least squares regression weights applied to all tests forming each test composite, the LSEs, provide maximum utility when used in either or both selection and classification. Such composites will not only provide the means of maximizing the average validities across jobs but will also maximize potential allocation efficiency (PAE). The validities of these composites are, of course, the multiple correlation coefficients between the composites and each job criterion measure. No set of composites selected to lower intercorrelations among composites or to increase the variation of composite validities across jobs (as one might mistakenly attempt to do in order to increase PAE) can increase the utility function value as well as the full regression equations based on the total battery. If composites use a reduced number of tests or otherwise are not LSEs, or if jobs are clustered rather than matched each with its own LSE, the best composites for selection are not necessarily the best for classification.

## C. IMPROVING CLASSIFICATION EFFICIENCY

The possibility of fully benefiting (i.e., maximizing allocation efficiency) with no decrease in average validity as a consequence, depends upon the following conditions:

(1) whether the battery and composites are already determined; (2) which optimal selection/classification procedure is being utilized to implement assignment to jobs (an LP type program); and (3) whether job families are appropriately structured (smaller differences among LSEs for jobs within families and larger differences between LSEs for jobs across families, or ideally, one LSE for each job).

The potential benefits of optimal assignment are usually not realized because of the nature of the operational assignment process used in practice. The traditional assignment approach used in the military, for example, is a two-stage process: selection is first accomplished based on AFQT entry-level recruitment standards, then classification is accomplished on the selected group through the use of aptitude area composites. Benefits, however, are maximized through the use of a single stage selection/assignment process (i.e., multidimensional screening, the MDS algorithm that integrates the effects of both selection and classification). Using the MDS model, both processes are accomplished simultaneously through the use of different cut scores optimized for each job family predictor composite. An optimal selection/classification process most probably has never been used in any operational context.

We define and describe means of defining and measuring potential allocation efficiency (PAE), potential classification efficiency (PCE) and potential utilization efficiency (PUE). The total selection, classification and placement process, individually or in combination, is termed the "personnel utilization decision process." As noted, classification efficiency may be subdivided into two effects: allocation efficiency and hierarchical classification efficiency. All classification efficiency not due to hierarchical layering effects, when heterogeneous validities and/or values are assigned to jobs and also reflected in the predictor variables used in the assignment process, is attributable to allocation efficiency. When the classification test battery is unidimensional, no allocation benefit can exist; the assignment process consists entirely of hierarchical classification. If all assignment variables (e.g., aptitude areas composites) have equal means and variances, the classification process is pure allocation since no means for an hierarchical classification process to capitalize on hierarchical layering is present. However, when hierarchical layering of validities or job values exists and is reflected in the predictors, and the joint predictor-criterion space is multidimensional, the classification process includes both hierarchical classification and allocation processes. When both hierarchical classification and allocation are present in the same process, their effects are so confounded as to make them difficult, if not impossible, to separate.

The work of Brogden and Horst generated the main stream of progress in the measurement and improvement of classification effectiveness. Their contributions are described in detail. Brogden's formulation ties classification to mean predicted performance (MPP) and thus to utility in dollar-valued terms. Horst's measure of classification effectiveness has a direct relationship to Brogden's measures when all of Brogden's assumptions are met. The square of Horst's index is proportional to Brogden's index, when all the assumptions of Brogden's 1959 model are met, and can be used to determine the rank order of alternative batteries in terms of PCE.

We describe methods of improving potential efficiency through test selection, job family restructuring, and/or selection and restructuring of test composites associated with various jobs. The use of factor analysis to examine test content and/or job clusters as they affect PAE or PCE is described.

A final topic discussed is the use of synthetic (generated) scores to simulate personnel utilization applications so that alternative policies and procedures may be fully evaluated without sampling distortions introduced by operational utilization of a battery for selection and assignment. Synthetic samples may be drawn to represent empirical data (e.g., test and criterion scores) and simulation studies conducted.

We assert that there is potential for more than three or four dimensions in the joint predictor-criterion space. Batteries developed to maximize selection efficiency and validated against limited, unidimensional job criteria are not the best starting point in finding additional dimensionality needed for classification efficiency. Finding more dimensionality in the joint predictor-criterion space requires at least the effort, concern and care that was used to confirm the existence of general mental ability, clerical speed, and psychomotor ability in the joint General Aptitude Test Battery-criterion space. The methodology suggested in this report is essential for identifying both the potential and existing operational utility of the ASVAB in classification.

# CHAPTER 1. PERSONNEL UTILIZATION IN THE ASSIGNMENT PROCESS

## A. INTRODUCTION

The central thesis of this report is that both personnel research and the operational implementation of research findings should have as their primary objective the improvement of utility. It is utility rather than the psychometric merit of the predictors that best justifies the use of tests to select, classify, and/or place personnel. In this report we extend this central theme from its emphasis on selection for a single job to selection and classification for multiple jobs, accomplished through reliance on the measurement of mean predicted performance (MPP). MPP is as useful a means of expressing benefits derived from classification and placement as it was shown to be in unidimensional selection.

While the literature on the utility of selection is rich, and growing rapidly, there has been very little published on the utility of classification and placement since Cronbach and Gleser (1965). Significant contributions to the methodology for measuring and improving classification benefits have been sparse since Paul Horst (1954-1960), Hubert Brogden (1946-1964) and Richard Sorenson (1965-1967) wrote about the impact of personnel classification on mean predicted performance.

We use the term "personnel utilization" to designate the total selection and assignment process. The effects of alternative personnel strategies, and of tactics that can manipulate variables in the personnel utilization system, can and should be measured in terms of MPP. We describe how to: (1) choose among tests for inclusion in operational batteries; (2) structure test composites and associated job families; and (3) design personnel utilization processes; all are discussed in terms of which alternative personnel strategy will best improve MPP.

Efficient utility analyses require consideration of all personnel utilization procedures; when present in the personnel system being analyzed, the effects of selection, classification, and placement must be reflected in the integrated computation of MPP. The utility of a personnel system is as much a consequence of the efficiency of the selection/assignment process as of the psychometric quality of the predictors.

An effective taxonomy can assist in the selection or design of efficient algorithms that can approach the full potential (theoretical) efficiency in actual operational situations. Also, a taxonomy with provision for: (1) selection procedures that utilize multi-dimensionality in the joint predictor-criterion space, (2) the application of selection and classification/placement procedures in either separate or combined stages, and (3) procedures that capitalize on the differences of the validities of assignment variables and/or the values of job performance across jobs, is essential to the measurement of both potential and operationally obtained utility.

Selection and classification is almost always linked sequentially with classification occurring after an initial selection process. We will refer to this model as the two-stage selection/classification process. There is a one-stage simultaneous selection-classification model that provides a feasible alternative to this two-stage model. The optimal simultaneous process for accomplishing selection and classification--the multidimensional screening algorithm (MDS)[1]--could be readily applied in the military setting, although thus far it has never been implemented operationally.

Personnel utilization processes divide into several categories, each having different implications for optimizing personnel utilization procedures and for measuring utility. A personnel utilization taxonomy is included as a means of providing precision in designating which process is under discussion. This taxonomy classifies personnel utilization processes into non-exclusive categories based on whether a process: (1) is unidimensional or multidimensional in the joint predictor/criterion space; (2) has the goal of rejecting applicants (selection); assigning accepted personnel for jobs (classification); or of assigning them to levels within jobs (placement); and (3) capitalizes on disparate validities or values across jobs.

Job levels are the rungs on a specific career ladder corresponding to a progression of skill levels; these rungs might, for some industrial jobs, be designated as trainee, apprentice, journeyman, or master positions. In the Army these job levels are the skill levels in a military occupational specialty (MOS), as 1 (the entry level) through 3 (the trainer/supervisory level). Placement into levels of a language or mathematics sequence separately from the selection process in the university setting provides an academic parallel to selection and placement for jobs. In distinguishing between jobs and job levels we are

1 This algorithm for the simultaneous and optimal accomplishment of selection and classification as an integrated process is described in Section C of Chapter 1.

thinking of jobs as military occupational specialties rather than the specific duty positions that are found within a MOS (many at the same level).

The "best" test composite for use in either selection or classification is a least squares estimate of performance (one LSE per job) based on all tests in the experimental test pool. A set of such composites is equally "best" for use in selection, classification, or placement. Any deviations from this ideal, including all further test selection to identify operational batteries or to form test composites, creates a requirement for separate consideration of selection and classification. Similarly, the clustering of jobs into families in order to reduce the number of operational test composites must be accomplished differently depending on whether the objective is to optimize selection, classification, or both.

A taxonomy of personnel utilization and a vocabulary in which potential efficiency measures are related to the primary utilization categories are provided in Chapter 1. We also emphasize the importance of the assignment algorithm's role in achieving the maximum benefits from either multidimensional selection or classification.

The contributions of Brogden and Horst to the measurement of classification efficiency are discussed in Chapter 2. The proportionality of the square of Brogden's measure of potential allocation efficiency (PAE) to Horst's index of differential validity ($H_d$) (if Brogden's assumptions are met and the number of jobs is held constant) is established in Chapter 2. The effects of hierarchical layering on $H_d$ is then discussed; when hierarchical layering is a major contributor to the magnitude of $H_d$, the lack of evidence for a close relationship between $H_d$ and MPP reinforces our preference for using MPP, the more direct measure of potential classification efficiency, instead of $H_d$ in the evaluation of alternative utilization strategies.

Brogden (1959) provides tabled values for mean orthogonal criterion scores. When Brogden's assumptions are met his entries can be multiplied by $R(1-r)^{1/2}$ to yield MPP standard scores where $R$ is the common multiple correlation coefficient for the LSEs and $r$ is the common intercorrelation among the LSEs. Brogden's model is of major importance because it proves that a classification process can be profitable even when the intercorrelations among LSEs are high. However, we emphasize that other predictors cannot be substituted for LSEs in Brogden's model. Further, we do not know how robust Brogden's and Horst's indices are as one departs from the assumptions of Brogden's 1959 model.

Relevant literature on the contribution of classification to utility is summarized in Chapter 2. Particular attention is given to studies that relate characteristics of the classification process to MPP. Studies and/or methodologies using subjective estimation of payoff as the figure of merit have been intentionally omitted from our review.

In Chapter 3 we proceed from the measurement of classification efficiency to its improvement through selecting predictors or structuring jobs and associated LSEs. A means of identifying rotated principal component factors that maximize $H_d$ is described and their use as composites or as a means of clustering jobs for use with composites is recommended.

Most attempts to improve the classification efficiency of the personnel utilization process will require the use of a readily computable index as a surrogate for MPP. While Horst's differential and absolute validity indices ($H_d$ and $H_a$, respectively) are used as figures of merit for most of the selection or clustering techniques described in Chapter 3, two theoretically superior indices are also proposed.

Dedication to the use of MPP as the figure of merit for evaluating personnel utilization strategies motivates our inclusion of a chapter on model sampling. We believe the use of a number of samples of real or synthetic data as input into simulations of personnel utilization strategies is the only practical way to obtai ı the MPP scores required for utility analyses. Chapter 4 is intended to help researchers and system analysts evaluate the validity and usefulness of model sampling for utility analyses and to provide a starting point for those who choose to use this methodology for computing MPP scores.

The complexity of multidimensional selection and assignment processes precludes the use of simple analytical methods for computing MPP scores required for utility analyses. This complexity contrasts with the simplicity of the univariate selection model in which the validity coefficient is directly proportional to MPP when the selection ratio is held constant and the relatively simple optimal selection algorithm is utilized (i.e., the rank ordering of applicants on predicted performance and selecting in order from the top down). When analytical techniques cannot provide the MPP scores in a metric compatible with the measures of cost obtained for a utility analysis, the remaining alternative is simulation.

The initial input for simulations designed to provide MPP scores may be either real records from a large data bank or entities consisting of synthetic scores provided by model sampling techniques. Because the availability of MPP scores is essential to credible utility analyses, we provide a description of model sampling techniques appropriate for use in simulating personnel utilization in a system context; the focus is on simulations designed to provide an MPP standard score as the final output.

The purpose here is not to provide a comprehensive treatise on classification. Actually, numerous topics in the classification domain have been deliberately left out because they are side issues in the context of the central thesis of this report. The content we selected for inclusion relates to the determination of: (1) the potential contribution of classification to utility, and (2) the utility provided by classification effects in an operational assignment process.

Classification is necessarily a multivariate topic and cannot be broken down into univariate terms without losing the essence of what is to be gained from the simultaneous consideration of many variables. Either relatively simple matrix notation or very complex and tedious summation notation must be used to express these multivariate relationships. We chose to use matrix algebra and to place the more formal derivations and demonstrations in the appendices.

Much use is made of a particular factor analysis solution--the Dwyer factor extension solution. The use of this factor solution ties together the contributions of Horst and Brogden and also provides insight into other proposed approaches. It is recognized that most readers will not claim to have more than a modicum of facility in matrix algebra, although it is anticipated that the majority will have some familiarity with the use of factor matrices to present results of psychometric studies. It is intended that the non-mathematical explanations in the text will carry the reader along even if he or she ignores the occasional use of matrix algebra.

## B. EXTENDING SELECTION UTILITY TO MORE COMPLEX DECISION SITUATIONS

In an earlier report selection utility was described and analyzed (Zeidner and Johnson, 1989a). In the report, we introduce the concept of classification effects as an ingredient of utility. Research publications on classification effects and utility are far fewer in number than on selection utility. Therefore, we start by making several distinctions in terminology between the well-known terms used in selection utility and those we must use to incorporate the effects of classification into selection and classification utility.

The term "personnel program effects in selection utility" refers to productivity gains attributable to the selection procedure based on net benefits (i.e., productivity gains minus program costs) expressed in dollar-valued terms. These gains can be referred to as "benefits." In this chapter we use the term "benefits" to embrace a set of procedures broader than selection alone. Additionally, we refer to performance benefits without consideration of program costs (e.g., costs of recruiting, testing and attrition). The utility of personnel program effects attributable to selection/classification procedures are considered in Zeidner and Johnson (1989b).

Personnel programs may result in benefits attributable to procedures not considered in previous chapters, including: (1) simultaneously selecting for several types of jobs; (2) selecting and placing applicants into an appropriate level of a job (as in the Army's "stripes for skills" program); (3) first selecting and then asssigning to the job for which predicted benefit is maximized; or (4) selecting, placing, and classifying personnel in separate stages or simultaneously as an integrated decision process.

Each of these procedures uses a distinct decision process that is used to select and/or assign personnel. We would expect each procedure to provide greater benefits (and utility) than is obtainable from a simple selection procedure. This potential increase in benefits distinguishes utility measurement that includes classification effects from utility measurement based on simple selection. Once the predicted benefits from classification have been determined, most of the concepts used in selection utility are applicable to classification. The determination of predicted benefit depends on: (1) the decision process itself; (2) the potential efficiency of the test battery; and (3) the dimensionality of the joint predictor-criterion space.

The disparate effects of selection, placement and classifiction on predicted performance requires a taxonomy which assists in the understanding of selection, classification and placement procedures, singly or combined, in the context of improving performance and measuring utility. Capabilities of various procedures for capitalizing on variances of predicted performance (PP) scores, between and within, people, jobs and job levels can be better understood in the context of this taxonomy.

This chapter provides precise definitions for selection classification and placement, the major procedures comprising a personnel utilization taxonomy. These major procedures are further broken down into subcategories based on whether or not they capitalize on: (1) multidimensionality in the joint predictor-criterion space, and (2) hierarchical value or validity relationships that link predictor and criterion variables. We also describe decision outcomes associated with these procedures: rejection versus acceptance; rejection versus assignment to specific jobs; and assignment of those already accepted. Certain decision processes can provide optimal outcomes for some procedural subcategories but not for others. However, the algorithms we call multidimensional screening (MDS) can, when used appropriately, maximize MPP for all three procedures and their subcategories.

Existing operational procedures and/or test batteries are often much less efficient than they could be. For example, the full set of tests in a battery often are not used in prediction equations or test composites. If least square estimates based on the total battery are not used, classification and/or placement procedures will not yield all of the potential classification efficiency (PCE) embedded in the battery. While simple validity of a best-weighted test composite (corrected for restrictions in range and criterion unreliability) is proportional to the potential selection effectiveness (PSE) of a specified set of tests, more complex procedures are required to estimate the corresponding PCE of a test battery. A PCE estimation must be made in the context of a specified set of jobs, job performance measures, test battery, test composite sets (assignment variables) and the assignment algorithm.

While the expression of PSE in terms of mean predicted performance (MPP) is optional, since PSE can also be expressed as a validity coefficient, PCE can only be expressed in terms of MPP. Thus, as a means of linking this publication with Zeidner and Johnson (1989a), we define both PSE and PCE as the MPP standard scores resulting, respectively, from selection or classification procedures.

As a starting point, in the determination of the contribution of classification effects to utility, the measurement of benefits can be approximated by computing mean predicted performance (MPP) across jobs. If MPP is weighted by the value (importance) of each job, it becomes a more useful measure of benefits. Thus the term "benefit" is used to denote a theoretically desirable measure of performance that is value weighted for jobs and/or job levels and is expressed in terms of an appropriate metric. This variable, when correctly combined with costs, provides a measure of utility.

The discussion of a personnel utilization taxonomy in the following section assumes that the goal of selection, classification, and placement procedures is, individually or in combination, to maximize mean predicted benefits. As mentioned earlier the total mean predicted benefit from a classification process is a function of the effectiveness of the selection/assignment algorithm (the decision process), potential classification efficiency of the battery, and the multidimensionality of the joint predictor-criterion space (the space spanned by the least square estimates of the multi-job criteria).

This chapter also considers the impact of policy constraints on the decision process. It is frequently necessary to make compromises between efforts to enforce constraints imposed to reflect personnel policies and efforts to maximize the benefit provided by the personnel utilization decision process.

The measurement or estimation of utility obtainable from using personnel instruments to make operational personnel decisions in the context of a selection process is covered in Zeidner and Johnson (1989a). An optimal selection process is frequently visualized in the discussion of utility estimation models as the rank ordering of all applicants on the benefit predicted from the personnel instrument(s), and the rejection of all those below a specified cut score on a predicted benefit continuum. Optimal processing for the more complex personnel utilization categories must also be similarly defined as the first step in determining their utility.

The literature bearing on the utility of personnel instruments has little to say on the benefits obtainable from: (1) simultaneously selecting for several types of jobs; (2) selecting and placing applicants into an appropriate level of a job (as in the Army's "stripes for skills" program); (3) first selecting and then assigning to the job which maximizes predicted benefit; (4) selecting, placing, and classifying personnel in separate stages, or simultaneously as an integrated decision process.

There are obviously many different ways, each using a distinct decision process, in which personnel instruments can be used to select and assign personnel. Many of these ways could provide greater benefits and a greater contribution to utility than is obtainable from a simple selection process. It is primarily the increment in the potential benefit obtainable from the personnel utilization process that makes it so important that classification be considered, along with selection, in the estimation of utility.

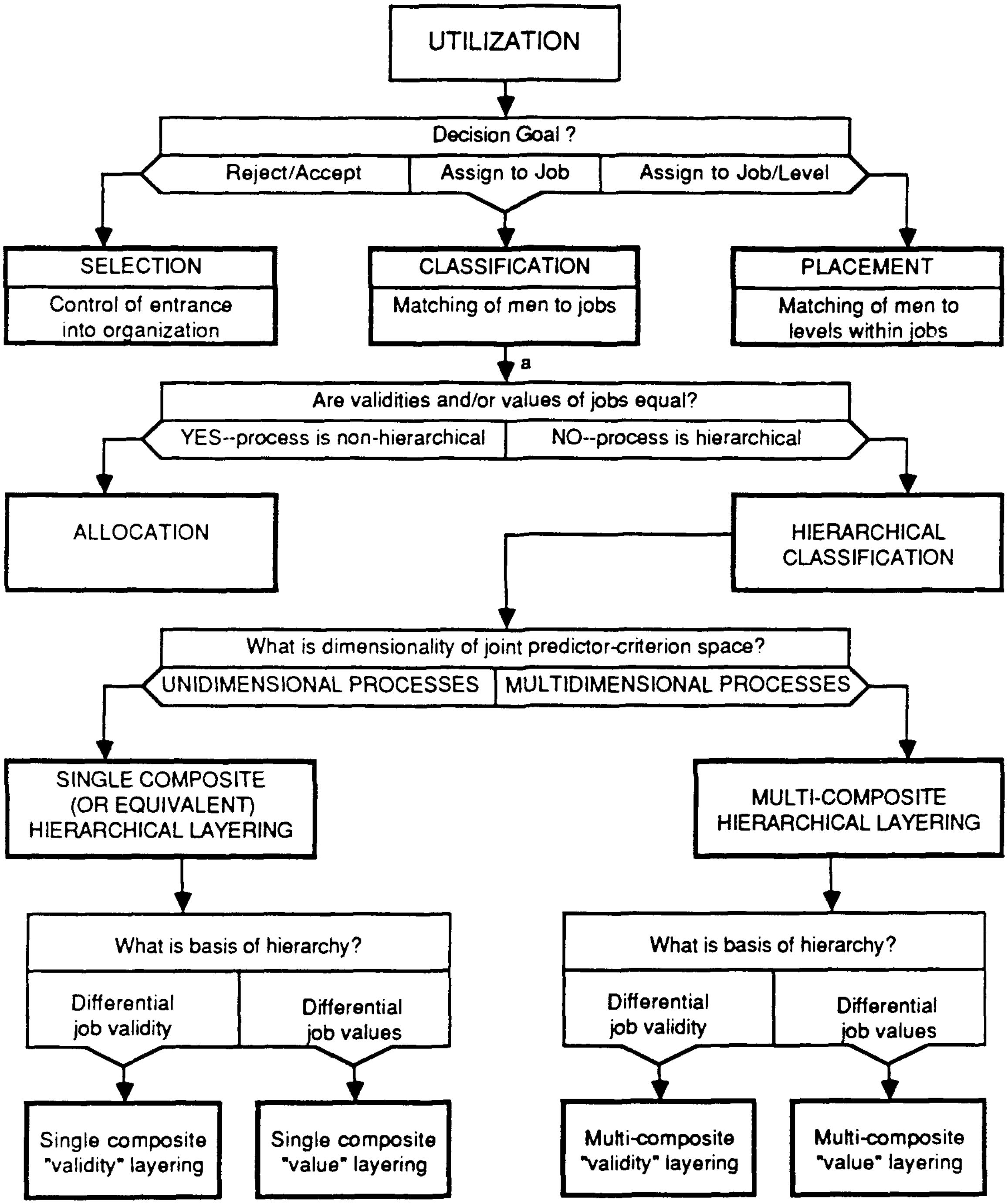
Once the predicted benefit has been determined, most of the utility concepts and estimation procedures discussed in Zeidner and Johnson (1989a) are applicable. The predicted benefit will be maximized when the following conditions are met: (1) The decision process for selecting and assigning is optimal; (2) the test battery and the test composites have been selected to maximize PAE, PCE, and/or potential utilization efficiency (PUE); and (3) the set of criteria which maximize the dimensionality of the joint predictor-criterion space is used to compute validities.

## C. A TAXONOMY OF APPLICANT/EMPLOYEE UTILIZATION PROCESSES

The magnitude of "benefit" resulting from personnel utilization depends heavily on the nature of the decision process, including the algorithms used, recruiting and counseling policies, and the characteristics of decision tools. Before discussing the relationship of the above process components to both potential and actual benefit, we first define a vocabulary of the processes that can be used to implement selection, classification, and placement policies. Definitions will follow common usage, except where there is not agreement on a precise meaning. Also, names and definitions are given to several processes that lack recognizable names, or have inconsistent definitions in the literature.

The total selection, classification and placement process, individually or in combination, is termed the "personnel utilization decision process." Utilization is subdivided into the three procedures or subcategories of selection, classification, and placement, in accordance with the goals and characteristics of the utilization process. (See Figure 1.1.) Each of these three subcategories is further subdivided into two sub-subcategories, each based on whether or not the decision process capitalizes on the mean predicted benefit scores variance, across jobs for classification, across levels within jobs for placement, or for either or both jobs or job levels for selection. A "hierarchical" process will be said to occur in selection, classification, or placement if the mean predicted benefit scores of those performing in different jobs or in different job levels are sufficiently different to make a practical difference, and the selection/assignment process capitalizes on these differences. (See Figure 1.2).

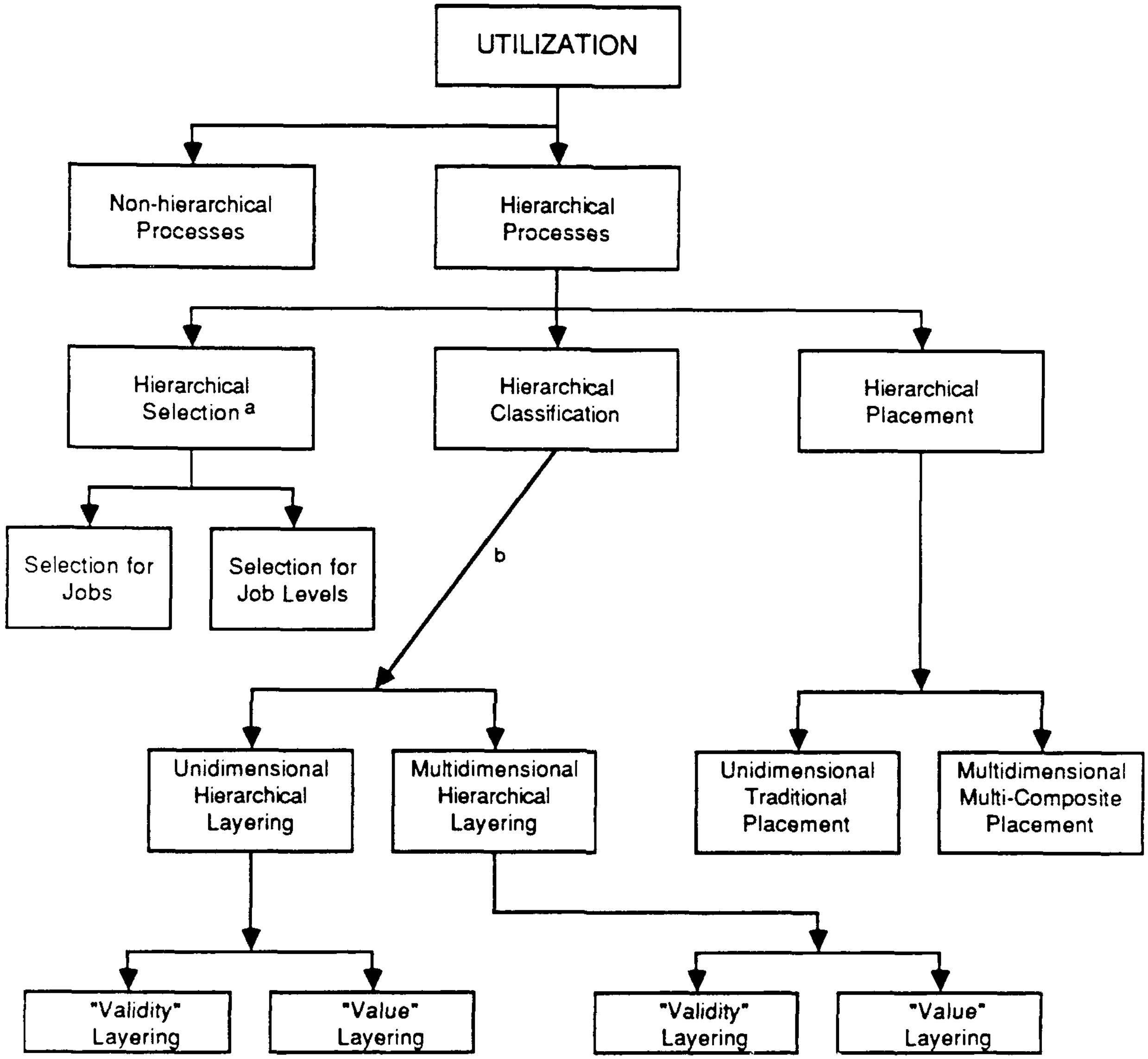
A difference among mean benefit scores across jobs (a hierarchical layering effect) can result from either differences in validities or in the differences in values (importance or criticality) accorded to jobs. Both differences may exist in the same situation. To capitalize on differences in validities, the most effective test composites will of course be the least squares predictions of benefits. Other test composites may not necessarily provide the maximum hierarchical layering capability, even though the other three conditions are present (i.e., the assignment or placement process is capable of capitalizing on the hierarchical effect and both the test battery and the performance measures have the characteristics that elicit the hierarchical effect). For example, the Army aptitude area composite predictors, using an optimal assignment algorithm, still would not provide a hierarchical layering capability, despite validities that vary considerably across jobs, since




NOTE:

[a] This branch could also be attached to either selection or placement. If attached to selection, "simple selection" is substituted for "allocation;" If attached to placement, "vertical job matching" is substituted for "allocation." The "Hierarchical Classification" branch becomes "Hierarchical Selection" if attached to "selection," and "Hierarchical Placement" if attached to "placement." Multidimensional Selection has an additional division into "selection for jobs" vs. "selection for levels within jobs."

**Figure 1.1. Taxonomy of Personnel Utilization Processes**




NOTE:

[a] Equal to Hierarchical Classification or Hierarchical Placement processes but with an added rejection category.

[b] This stem also attaches to each of the two divisions of Hierarchical Selection, and could substitute for the stem shown attached to Hierarchical Placement.

**Figure 1.2. An Alternative Depiction of Personnel Utilization (an Emphasis on the Hierarchical Processes)**

they were standardized to have equal means and variances. However, if the composites were to be weighted by job values or by the validity of each composite for a job, and assignments accomplished using an optimal assignment process, a hierarchical layering effect could result.

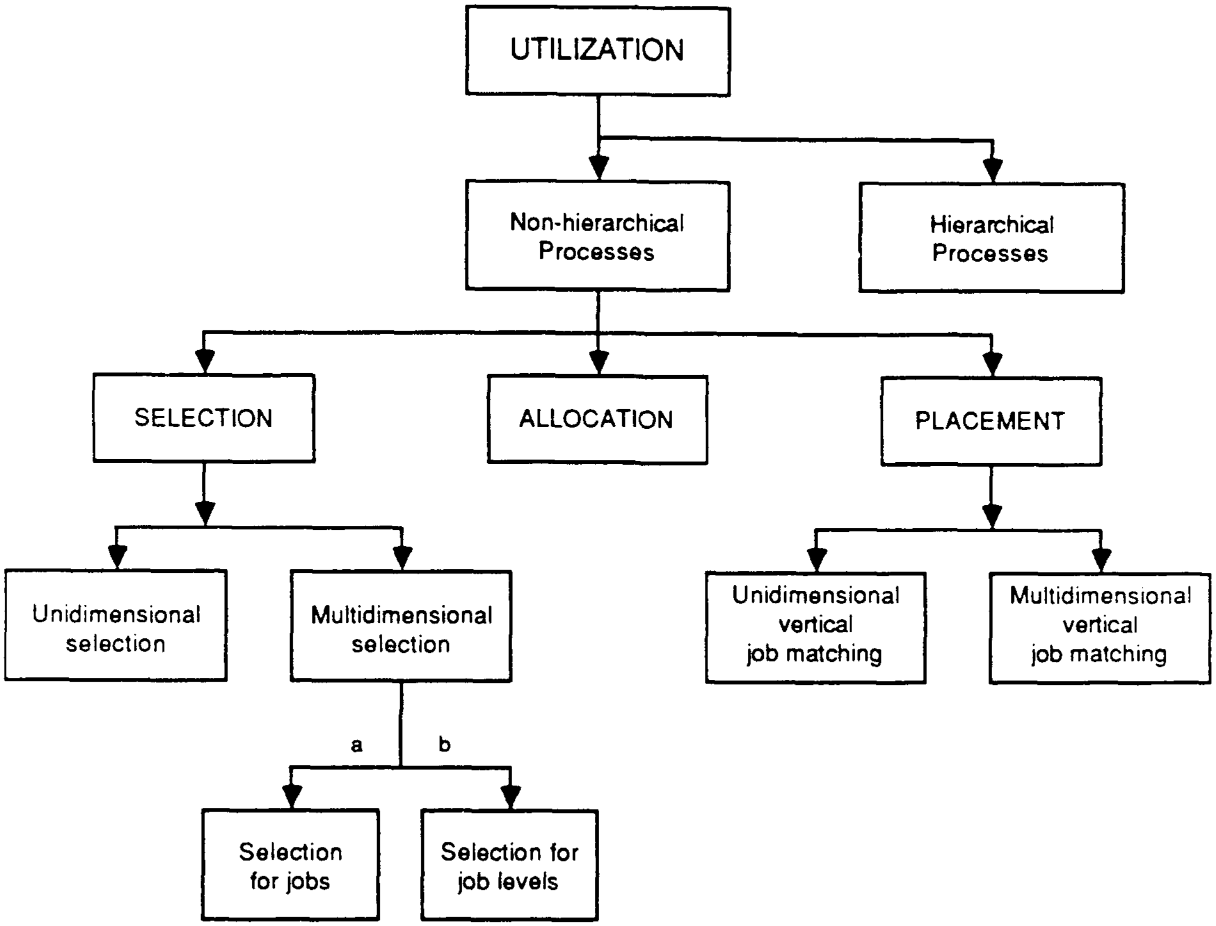
The absence of a hierarchical capability identifies the process as being in one or more of the following three sub-categories: (1) traditional selection, (2) the allocation subcategory of classification (a horizontal job matching process), and/or (3) a special kind of placement within upper and lower cut scores which we refer to as the vertical matching of personnel skill levels to job requirements. (See Figure 1.3.)

### 1. The Six Process and Three Tool/Data Characteristics

The subcategories are distinguished from each other by the presence or absence of six decision processes and three tool/data characteristics. (See Tables 1.1 and 1.2.) The first process characteristic, PC1, relates to the result or goal of the decision process. The three major categories are uniquely defined by the decision goals (i.e., to accept or reject applicants is a selection goal, to assign across jobs is a classification goal, and to assign across job levels is a placement goal). Additionally, the major categories also have distinct relationships to the remaining five process characteristics and could be uniquely defined by the presence or absence of characteristics critical to each category, entirely apart from PC1.

The second decision process characteristic, PC2, is essential to a selection process. The selection algorithm must have the capability to rank order applicants on a predicted benefit continuum in order to accept those yielding the highest mean predicted benefit. In other words, the selection process must have the capability of capitalizing on the spread or variance of predicted benefit scores among the applicants for each of one or more jobs.

The third decision process characteristic, PC3, relates to the capability of the decision algorithm to capitalize on intra-individual differences, the variance of predicted benefit within each individual across jobs. The presence of this capability is a necessary and sufficient characteristic of the classification subcategory we refer to as allocation.




NOTE:

[a] Multidimensional classification process with a rejection category at the lower end of each continuum.

[b] Multidimensional placement process with a rejection category at the lower end of each continuum.

**Figure 1.3. An Alternative Depiction of Personnel Utilization (an Emphasis on the Non-Hierarchical Processes)**

Table 1.1. The Personnel Utilization Decision Processes

| One Stage Procedures | PC[a] | Objective | Optimal Process |
|---|---|---|---|
| Selection | 2 | Reject/accept | Selection = rank order and determination of cutting score on continuum |
| Allocation | 3 | Assign to job | Job assignment = LP algorithm for accomplishing person/job match |
| Vertical Job Matching | 4 | Assign to job level | Job level assignment = LP algorithm for accomplishing person/job-level match |
| Hierarchical Classification | 5 | Assign to job | Job assignment[b] |
| Hierarchical Placement | 6 | Assign to job level | Job level assignment[b] |
| Selection-Classification/ Placement | 2, 3, 4, 5 | Reject vs. assign to job and/or job level | Selection and job and/or job level assignment[b] |

NOTE:

[a] See text for definition of process characteristics (PC).

[b] Optimal processes for selection and assignment are defined above.

**Table 1.2. Relationship of Selected Procedures to Variance of Benefit Measures**

| Identification of Procedure | Tool Data Characteristics | Objectives | Identification of Variance Required for Process Efficiency |
|---|---|---|---|
| Traditional Simple Selection | 1a | Reject/accept | Within a job, across individuals |
| Simple Placement-Selection | 1b | Reject/accept | Within each job level, across individuals |
| Allocation | 2a | Assign | Within each individual, across jobs |
| Vertical Job Matching | 2b | Place | Within each individual, across job levels |
| Hierarchical Classification | 3a | Assign | Job means across jobs |
| Hierarchical Placement | 3b | Place | Job level means across job levels |
| Hierarchical Job Selection | 1a and 2a and/or 3a | Reject/assign | Within a job means across individuals and within each individual, across jobs, and/or job means across jobs |
| Classification | 2a and/or 3a | Assign | Within each individual, across jobs and/or job means across jobs |
| Placement | 2b and/or 3b | Place | Within each individual across job levels and/or job level means across job levels |
| Horizontal Job Selection | 1a and 2a | Reject/assign | Within each job across individuals and, within each individual, across jobs |
| Vertical Job Selection | 1a and 3a | Reject/place | Within each job level, across individuals and, within each individual, across job levels |
| Multidimensional Selection (relating to jobs and/or job levels) | 1a and one or more: 2a, 2b, 3a, and/or 3b | Reject/assign | Within each job and/or job level, across individuals and at least one of the following: (a) within each individual, across jobs (b) within each individual, across job levels (c) job means across jobs (d) job level means across job levels |

The fourth decision process characteristic, PC4, similarly relates to the capability of the decision algorithm to capitalize on the variance of predicted benefit provided by each individual across layers or levels within one or more jobs. Just as the presence of PC3 defines allocation, PC4 defines the placement sub-subcategory that does not capitalize on hierarchical layering. We name this subcategory vertical job matching. This is not traditional placement in which each applicant is assigned to the highest level for which he or she qualifies, but instead gives equal emphasis on not placing an individual at a level where he or she would perform more poorly as a result of being over qualified (and perhaps under motivated). This process was implemented in the first linear program (LP) driven assignment program utilized by the Marine Corps; the Air Force utilized a similar concept in their enlisted classification program.

The fifth process decision characteristic, PC5, pertains to the capability of the decision algorithm to capitalize on the variance of mean predicted benefit scores across jobs. This characteristic is essential for hierarchical layering, a subcategory within classification. As previously noted, this variance can be the result of either different values placed on comparable performance for different jobs, variance in validities across jobs, or both. For this capability to be maximized, the test composites used in the classification process must reflect the values and/or the validities attached to jobs (i.e., have predicted performance means proportional to job values and/or validities).

The sixth decision process characteristic, PC6, similarly pertains to the capability of the decision algorithm to capitalize on the variance of mean predicted benefit, but across levels within jobs, rather than across jobs. This characteristic is essential for the hierarchical placement subcategory. This subcategory is the traditional placement process in which the employee or student is assigned to the most difficult tasks that the tests predict the individual is competent to perform. It is usually assumed that the benefit to the individual and/or to the organization is greatest when the individual is assigned to the most complex task he or she is competent to perform.

The three tool/data characteristics represent characteristics that must be present in the data for a process to effectively select, classify, or place personnel in jobs and/or levels within jobs. The first characteristic, TC1, relates to selection. Effective selection requires an adequate variance of predicted benefit scores for the target job.

The second characteristic, TC2, relates to both non-hierarchical classification and placement. Effective vertical placement and effective allocation require an adequate variance of predicted benefit scores within each individual--across jobs for allocation and across levels within jobs for vertical job matching (non-hierarchical placement). This tool/data characteristic is also required for effective multidimensional selection, as when multidimensional screening (MDS) or some other multidimensional selection algorithm is utilized.

The third characteristic, TC3, also relates primarily to classification or placement, but as with TC2 can affect simultaneous selection-classification or selection-placement. Effective hierarchical-classification and hierarchical-placement require an adequate variance of mean benefit scores, across jobs for hierarchical-classification and across levels within jobs for hierarchical-placement. The presence of the second of the above tool/data characteristics, TC2, is essential to effective allocation or vertical job matching (non-hierarchical placement) and is the result of predictor variables and job performance measures capable of combining to yield a multidimensional joint predictor-criterion space.

While PC1 places each process in an exclusive category, it should be obvious that neither the other five processes, nor the three tool/data characteristics, that together define and limit the subcategories of selection, classification and placement, are mutually exclusive. All categories can be effectively present in a single integrated process. For example, selection, allocation, and hierarchical layering could be accomplished in a single stage decision process in which each applicant is accepted or rejected; those applicants not rejected are assigned to jobs and/or to alternative levels within those jobs by means of a predictor battery, a decision process algorithm, and criterion information that can capitalize on all three of the tool/data characteristics.

We used the three best understood and universally recognized procedures--selection, classification, and placement--as our first level division in Figure 1.1. Utilization can be divided into these three categories on the basis of the goal or objective of the procedure (PC1); the same identification of a process such as selection, classification or placement, can be made by reference to the kind of benefit variance used by the process to make selection and/or assignment decisions (PC2, PC3, PC4, respectively).

Structuring the personnel utilization process along the lines described in our taxonomy is significant for a number of reasons. First, predicted benefit may be estimated differently depending on the process subcategory. Second, different test battery characteristics are desirable depending on the process being utilized. Third, assignment algorithms will be more efficient with respect to some utilization subcategories than to others. Fourth, and possibly most importantly, because different subcategories of utilization procedures lead to different distributions of "high quality" personnel assigned to critical jobs, the effects of differences in procedures impact on manpower policy. Alternatively, manpower policy directives affect the procedure subcategories in different ways and to different degrees.

## 2. Selection

In developing a taxonomy of utilization, one runs into an immediate problem concerning the appropriate dividing line between selection and classification. Others have determined the dividing line by differentiating between: (1) the filling of one job (selection) or more than one job (classification); or (2) rejecting some applicants (selection), or assigning all applicants to jobs (classification).

Our taxonomy is based on the latter distinction. (See Figure 1.1.) We identify selection as the procedure that produces the decision as to whether or not an individual becomes a member of the organization. A process is selection if the applicant is being considered for membership at large (with assignment to a job to come later) or if the applicant is being considered for rejection or acceptance for a number of specific jobs. Thus, selection can be either a unidimensional or a multidimensional process--selection for a single job or for many. In the latter case the amount of differential validity in a selection battery is an important determiner of the benefit resulting from a selection process.

The unidimensional selection procedure divides into: (1) the traditional selection process in which an applicant is accepted or rejected for a single job (or for membership in an organization); (2) the placement-selection process in which the individual is either rejected (non-selected) or placed at alternative levels in a single job (or family of jobs); and (3) the hierarchical classification-selection process in which some are rejected and the remainder assigned on the basis of hierarchical layering.

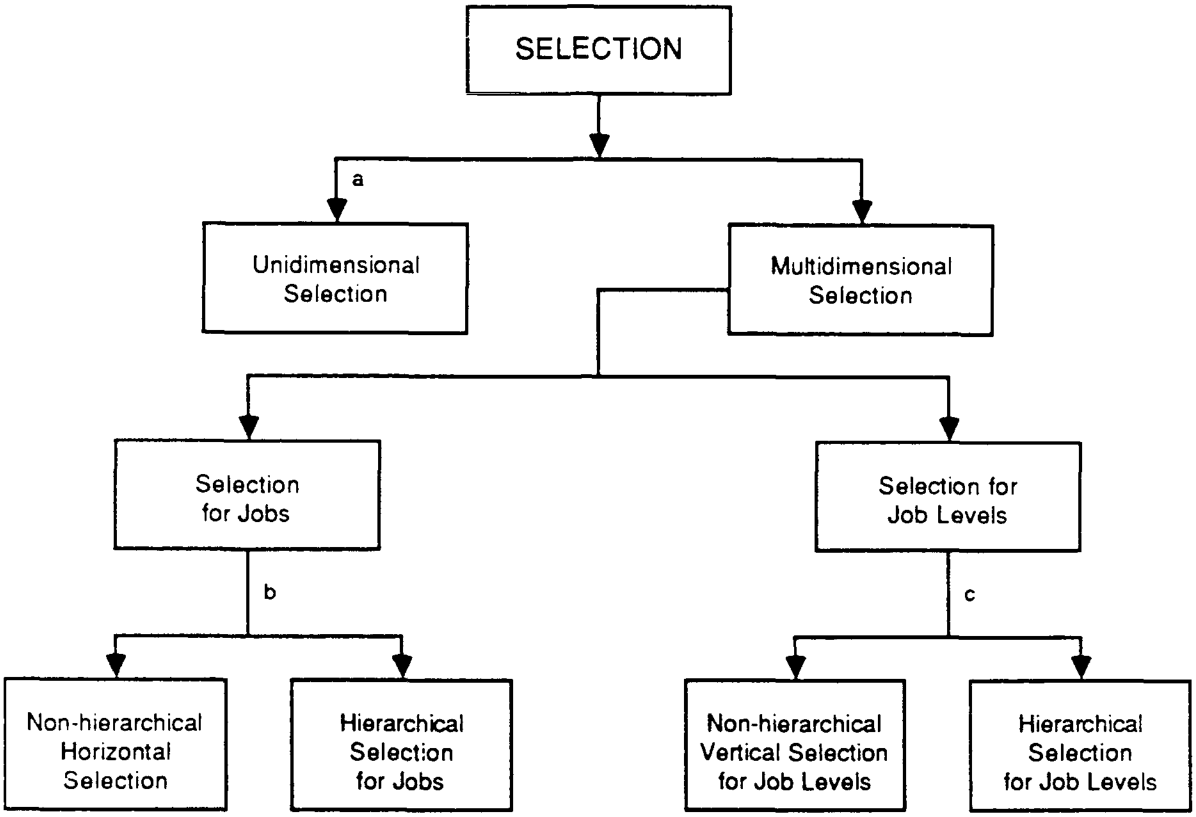
When there is only one selection instrument and several jobs of different values to be filled, predicted benefit may be maximized by rank ordering both applicants and jobs; the highest scoring applicant is assigned to the highest valued job, and assignment continues from the top scoring applicant downward with applicants scoring below some point rejected. The assignment process and the method for determining benefit is very similar in selection for hierarchical layered jobs to that of unidimensional selection for hierarchical placement. Both procedures are forms of simple unidimensional selection integrated with, respectively, hierarchical classification or hierarchical placement.

The multidimensional selection subcategory divides into selection for jobs or for job levels, each of which in turn divides into: (1) a subcategory where each job is equally valued, called here horizontal-selection (for jobs), or vertical selection(for job levels); and (2) a subcategory that capitalizes on the different values or validities of jobs. If a rejection category is provided for any classification or placement process, that process becomes selection by our definition (PC1). All optimal selection processes involve rank ordering applicants on some benefit continuum and rejecting all those below some point on that continuum. In multidimensional selection there is at least one more such continuum than for unidimensional selection; however, once each continuum has been created the decision process is essentially the same. (See Figure 1.4).

To achieve the maximum possible benefit out of simultaneous selection for a number of jobs, apart from the hierarchical phenomenon, the predictor battery must have what we refer to as potential allocation efficiency (PAE). For PAE to be non-zero, a multidimensional joint predictor-criterion space must exist. The operational assignment algorithm must capitalize on this potential if the operational allocation efficiency (OAE) is also to be non-zero. An effective assignment algorithm for maximizing the benefit of the selection/classification process should ensure that no nonselected person has a higher predicted performance on any job than the person assigned to that job. The algorithm should also ensure that no other assignment method can raise the mean predicted performance (MPP) further. We call one such algorithm that accomplishes both selection and classification, simultaneously and optimally, multidimensional screening (MDS), and describe it in a later section.

### 3. Classification

Classification is defined as the procedure in which employees are matched with jobs. The objective is to maximize the mean predicted performance (MPP) of those assigned. In a simultaneous selection/classification process, selection refers to the rejection or acceptance of applicants; classification relates to matching jobs and employees. Since the process is integrated, it may not be possible to say whether a given step belongs to either the selection or the classification aspect of the algorithm. Although the selection objective is usually stated in terms of maximizing the MPP of the selected group, fairness of the selection process is usually weighed in terms of the relative merits of individuals in the rejected group. No selection process can be said to be completely fair as long as a




NOTE:

[a] Traditional simple selection falls within this branch.

[b] Comparable to Classification but with a rejection category for each continuum.

[c] Comparable to Placement but with a rejection category for each continuum.

**Figure 1.4. An Alternative Depiction of Selection (an Emphasis on Multidimensional Selection)**

single member of the rejected group has a larger MPP for any job for which they are an applicant than the lowest scoring member of the accepted group. Thus the selection objective may be stated as that of minimizing the predicted benefit score of each applicant in the rejected group with respect to the specific job for which he or she comes the closest to being accepted. Since doing this will ensure that no rejected applicant is more qualified on any job than those accepted and assigned to that job, fairness or merit and utility are both served. The classification objective may also be stated as that of maximizing the mean predicted benefit score of the assigned group of employees.

Classification may be further divided into two processes, allocation and hierarchical classification. The allocation process capitalizes on the variance of predicted performance within an individual sometimes referred to as differential validity. Classification that is accomplished without capitalizing on hierarchical layering is "pure" allocation. A pure allocation process can be implemented in an optimal assignment process only when the criterion variables for each job or job family have equal validities and values (importance or criticality). We call the gain in benefit over random assignment obtained from this process "allocation efficiency," and the maximum effectiveness achievable from a given test battery and set of jobs, expressed as an MPP standard score, will be called potential allocation efficiency, or PAE. We should immediately note that PAE may be zero for the above example if the battery lacks differential validity or the criteria are unidimensional, even though the assignment process may be optimal.

All classification efficiency not explainable as allocation efficiency is attributed to hierarchical classification efficiency. Hierarchical classification is that part of the classification process that capitalizes on the disparate means and/or variance across the criteria. Even when the absence of differential validity prevents allocation effects, hierarchical layering can provide classification efficiency. In such a situation hierarchical classification efficiency can be demonstrated from the placement of each person in rank order on a predicted benefit continuum, using one predicted benefit continuum for each job, and entering each individual on a continuum as many times as there are jobs. Starting at the top of each continuum and proceeding downwards, the individuals are then placed in a job corresponding to the rank order of each score until the quota is met for a job. In progressing down each continuum, the scores for filled jobs are skipped over. Thus in a multi-job situation, pure hierarchical classification (i.e., no allocation effects are present) becomes almost indistinguishable from the "placement" procedure for one job. That is, a hierarchical classification process becomes computationally equivalent to a hierarchical placement process (traditional placement) as the joint predictor-criterion space approaches unidimensionality.

Hierarchical classification subdivides into a unidimensional and multidimensional category, just as is the case for the selection procedure. In turn, each of these categories subdivides into two subcategories based on how the hierarchy is determined. One approach capitalizes on the hierarchy of predictability of jobs, using a process that assigns individuals to jobs using job predictors (assignment variables) having variances proportional to their validities for each job. The allocation sum of mean predicted benefit scores is at a maximum when the predictor scores used in the assignment process are also least square estimates of benefit. Thus there is an obvious advantage to using least squares estimates of benefit for operational test composites. Such estimates have a variance equal to the square of the multiple correlation coefficient of the estimate with the criterion and can be expected to vary across jobs.

An assignment process using least squares estimates based on the full test battery as the source of operational test composites used to make assignments thus, is partly hierarchical classification--unless the estimates are standardized to have equal means and validities (e.g., when least square estimates in standard score form are divided by a number proportional to their validities to give them equal variances) prior to their use with the assignment algorithm.

The other method (subcategory) of hierarchical classification uses multipliers of the performance estimates of individuals corresponding to those values management (or some other authoritative source) places on performance in each job, to arrive at predicted benefit scores.

The designation of the classification process as being either allocation or hierarchical classification is straightforward and precise only under special conditions. For example, when no hierarchical layering effects exist, all existing classification efficiency is due to allocation efficiency. When there is only one predictor composite used for both selection and classification, all classification efficiency is due to hierarchical classification. However, when the joint predictor-criterion space is multidimensional and hierarchical layering is also present, the separation of classification effects becomes difficult and essentially ambiguous unless simplifying assumptions are made. Such a set of simplifying assumptions is made in Appendix 1B for the four variable model and in Chapter 3 to separate the contributions of hierarchical layering and allocations effects to an index of differential validity.

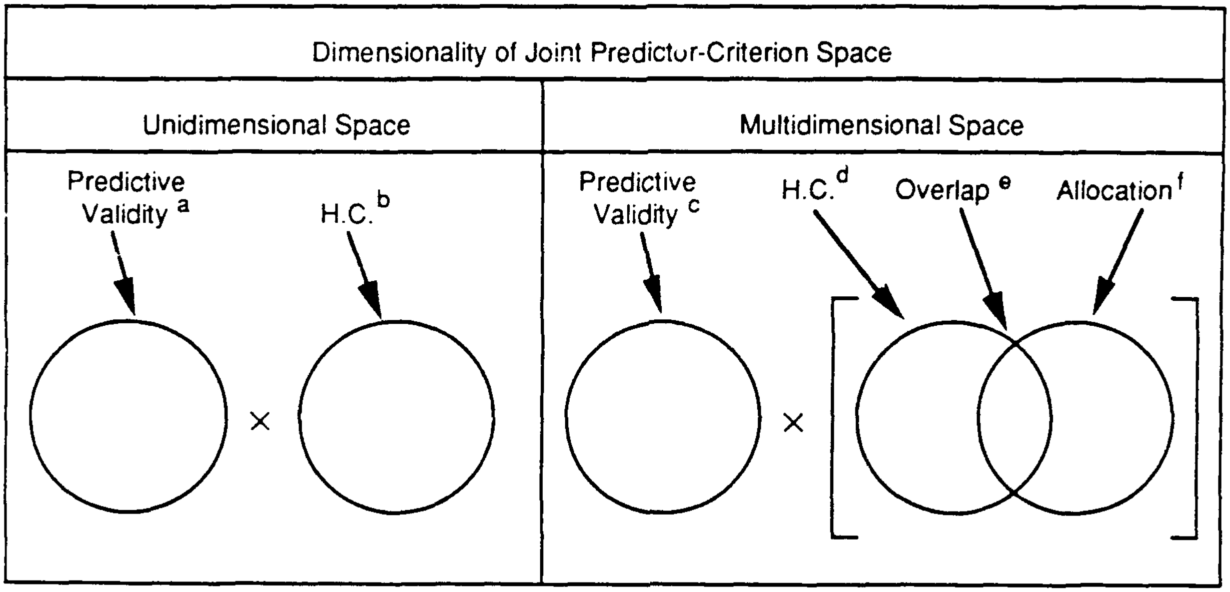
Hierarchical classification processes can be further divided into: (1) the unidimensional hierarchical layering case, (2) the multidimensional classification case where hierarchical layering effects are not duplicated by allocation effects, and (3) hierarchical layering effects that compete and thus, in effect, overlap (are redundant) with allocation effects. This third category is the portion of hierarchical layering effect which does not make an additional contribution to PCE over that provided by the within-person variance of the PP scores. These three categories of hierarchical classification, plus allocation, are depicted in Figure 1.5. Predicted validity is depicted as having a multiplier effect on allocation and HC, individually or together.

To profit from any category of hierarchical classification processes, the assignment variables must reflect the means and/or variances of the criterion variables. Thus, the existing Army aptitude areas, which are composites that have been standardized to have means of 100 and standard deviations of 20, cannot capitalize on this aspect of hierarchical classification.

An LP algorithm will provide optimal assignment for any of the various categories of classification; however, more simple algorithms will also provide optimal assignment when a hierarchical classification or placement process is unidimensional. Such a "more simple algorithm" will be illustrated in an example provided in a later section of this chapter.

### 4. Placement

Placement, by our technical definition, is analogous to classification. If we replace "jobs" with "levels within a job" in the definition of classification, we almost arrive at a definition for placement. In the placement procedure, individuals are matched to levels within jobs as compared to the classification process of matching personnel to jobs. As with classification, there is a subcategory of placement that capitalizes on a hierarchy of mean predicted benefits; this subcategory is called hierarchical placement. Similarly, there is an alternative subcategory (the non-hierarchical case) that may reasonably have been named allocation-placement. Instead, we call the latter subcategory, of vertical matching of individuals to job levels, "vertical job matching." It should be noted that the comparable subcategory of classification we call allocation may have reasonably been called "horizontal job-matching," except for the widespread use of the term "allocation".




a Predictive validity does not by itself provide PCE, but must be non-zero for some jobs if H.C. is to be non-zero; the average predictor validity across all jobs could be zero, providing a PSE of zero, and still permit the H.C. effect to be of considerable magnitude since it is the separate validities for each job, not the average validity, which provide the multiplier effect.

b An increment in MPP can be obtained from the use of separate assignment variables for each job or job family that reflect the disparate means and/or variances of the criterion variables; in this case the H.C. effects are based on a single predictor variable converted into predicted performance measures that match a continuum of predicted benefits.

c Some investigators appear to believe that the contributions of multidimensionality to PCE is entirely due to an increase in predictive validity; in fact, an increase in predictive validity due to use of separate LSEs for each job or job family, is one factor, but not necessarily the most important one, in providing the gains in PCE that often results from an increase in dimensionality of the joint predictor-criterion space.

d Hierarchical classification effects result from the matching of a hierarchy of predicted benefits, layer by layer, with corresponding layers of the assignment variable that have been rank ordered on a predicted benefits continuum.

e This overlap represents PCE that can be provided by either H.C. or allocation efficiency. Total classification effects are provided by the union of H.C. and allocation effects, not by their sum. Most of contribution that a moderate amount of H.C. can make to MPP when no allocation effects are present can be provided by a moderate amount of allocation efficiency in the absence of H.C. efficiency. Examples are provided in Appendix 1B.

f Allocation is the contribution of within person variance to PCE.

**Figure 1.5. Relationship of Hierarchical Classification and Allocation.**

As in selection, each of the two second-level subcategories of placement subdivides into a unidimensional and a multidimensional third-level subcategory. This dimensionality pertains to the joint predictor-criterion space, discussed in greater detail in the next two chapters. Obviously, the presence of only one predictor variable in the assignment (i.e., placement) process makes unidimensionality inevitable, but unidimensionality may result even with the use of multiple predictors. If only one valid factor (e.g., the "general mental ability" factor) is present in the predictor-criterion space for placement purposes, possibly because the criteria (the set of predicted benefits) are unidimensional, the "assignment" process would be within the unidimensional subcategory. Multidimensionality of placement predictors can be used across jobs (e.g., when the assignment variables are differentially valid across jobs for vertical job-matching or for hierarchical placement within each job). Multidimensionality could also be applicable for a single job when a separate predicted benefit (assignment variable) is computed for each job level within a job.

In providing examples of multidimensional placement, it is difficult to differentiate among placement subcategories since the nature of the relationships across jobs are not usually specified; cut scores on predictors, rather than optimal assignment algorithms with well defined objective functions are usually utilized to effect placement. Also, simultaneous consideration of examinees for several alternative levels across multiple jobs is probably a rarity in practice.

Advanced placement tests administered to entering college students to determine eligibility for receiving course credit (e.g., in calculus, or in a foreign language) are familiar uses in the academic setting. The Army's "stripes for skills" program and the Navy's World War II Seabee program that permitted experienced construction foremen to enter at senior petty officer grades are examples of placement in the military. The utilization of redundant employees in government and industry usually involves a placement process. The determination of an applicant's appropriate grade level for a civil service job on the basis of an unassembled examination is another example of placement. The use of placement is more common than the number of research efforts undertaken to evaluate utility attributable to placement procedures would lead us to believe.

Although our focus is primarily on classification and secondarily on multidimensional simple selection, placement is included as a procedure in this taxonomy in order to reduce possible confusion of hierarchical classification and placement and also to provide a complete taxonomy. Fuller consideration of the utility of placement for use in educational and employment contexts is worthy of separate treatment in future publications.

Placement has been given many different definitions in the literature. We are concerned with these definitions primarily because we wish to avoid confusion of placement with classification. Placement is a distinct process, not just a special case of classification when only one measure is being used to assign personnel to jobs. We believe it is important to distinguish between classification and placement, and to be able to use a terminology that permits the consideration of both unidimensional and multidimensional test and criterion sets for all three major processes: selection, classification, and placement.

Placement is defined by Cronbach and Gleser (1965) as the assigning of an individual to a "treatment" level using possibly one, but preferably more, test composites in the decision process. This definition is consistent with our taxonomy when "treatment" is restricted to personnel assignment.

Placement is defined by Schmidt (1988) as the assignment to one of several job alternatives when there is only one test composite, e.g., a general cognitive aptitude measure. His "placement" process is equivalent to our unidimensional hierarchical classification process.

Anastasi defined placement in her book "Psychological Testing" (1988) in terms of assignment to levels within jobs or training programs. Although she states, "...assignments are based on a single score" (p. 189), in describing placement, it is clear that she would restrict the use of the term placement to refer to the making of personnel assignment decisions with respect to a single job, where "...it is evident that...only one criterion is employed, and that placement is determined by the individual's position along a single predictor scale...although placement can be done with either one or more predictors, classification requires a multiple predictor whose validity is individually determined against each criterion." (p. 189). We accept her distinction between the focusing on one job or multiple job criteria as the basis of distinguishing between placement and classification. However, we extend both concepts on the predictor side to include both unidimensional and multidimensional processes. Both placement and classification can be based on use of either a single measure or a set of composites to make decisions about matching persons to jobs, or job levels.

Cronbach and Gleser (Psychological Tests and Personnel Decisions, 1965) utilize the term placement in a manner entirely consistent with our definition when they are referring to personnel procedures used in making assignments to levels of responsibility, to compensation levels within a job, or to difficulty levels in a training program (p. 54).

However, they extend their definition of placement to clinical diagnosis, and to the selection of alternative treatment of individuals in many situations, including the paroling of prisoners. No examples of personnel classification across jobs, defined as above, is included as a "treatment" in a placement process.

The desirability of considering the differential validity of predictors in the selecting of test composites to be used to effect placement to alternative treatments is emphasized by Cronbach and Gleser (1965), "A measure that predicts success under one treatment and not the other would be a much better aid to placement than a measure that predicts both" (p. 59). It is clear that both classification and placement measures are most efficient when a set of test composites possessing differential validity are available, in contrast to the use of a single general measure.

A tree structure is not needed to depict our taxonomy; the utilization of the particular inverted stem-to-leaf trees in the figures of this chapter have been provided to aid visualization of the taxonomy. However, the inherent structure of our taxonomy can be shown by using any of the major division principles as the first branching level. Figure 1.2, for example, illustrates that the first branching level could just as well be a binary one of hierarchical versus non-hierarchical instead of the triad of selection, classification, and placement. Or alternatively, this first branching level could be unidimensional versus multidimensional. The outcome for the final subdivisions would be the same. It should be noted that the division of the hierarchical processes into validity or value hierarchies (both can be present in the same procedure) is unique to these processes and cannot be extended to the non-hierarchical processes.

## D. THE ROLE OF MEAN PREDICTED PERFORMANCE AS A UNIFYING MEASURE READILY CONVERTIBLE TO UTILITY

The primary purpose of this section is to compare traditional selection with both multidimensional selection and classification with respect to the manner in which benefit and predicted benefit are defined and measured. The use of mean predicted performance (MPP) as a surrogate of mean predicted benefit indicates that MPP is the variable to be maximized in selection and classification processes. The substitution of MPP for mean predicted benefit is justified, inasmuch as we believe MPP is the common thread that links selection, classification, and placement on the benefits side of utility formulations. A number of utilization efficiency concepts used extensively throughout the remainder of this report will be defined.

We distinguish between operational efficiency and potential efficiency of personnel utilization processes. The measurement of personnel utilization efficiency, either for the actual operational process or for estimating the potential of a test battery, requires appropriately defined and computed scores for both the variables used to select and/or to assign and the scores used to provide the estimate of efficiency. Operational efficiency is the improvement in MPP obtained from the usually imperfect operational assignment process; potential efficiency is the improvement that would be obtainable if the maximally efficient prediction composites of a given battery were to be used in optimal selection/assignment algorithms. The resulting improvement must be measured in terms of the best obtainable "least squares estimate" of performance. We refer to this best estimate as predicted performance, and the measures of process efficiency will be expressed as an MPP standard score. The "process" for which efficiency is determined includes the selection/assignment algorithms, the battery, the choice of assignment variables, the set of jobs, and the performance measures.

The use of MPP as a measure of potential efficiency provides a means of comparing the effectiveness of alternative tests or test batteries in the context of a specified set of jobs and performance scores. Also, the benefit obtainable from an experimental pool of tests, using various combinations of selection, classification, and placement is expressible in terms of a measure of potential utilization efficiency. We later define and use measures of potential utilization efficiency (PUE), potential selection efficiency (PSE), potential allocation efficiency (PAE), and potential classification efficiency (PCE).

Potential efficiency measures for a specified test or test battery must use least square estimates of performance for both the variables used in the selection/assignment process and for the variables used in computing the MPP standard score for selected and assigned personnel. However, the assignment variables are estimates based on the specified test or test battery, while the estimates of performance used to compute the final result, the MPP standard score for selected and assigned personnel, are based on all available information, including all tests in the experimental battery and any other biographical or operational effectiveness variables for which the necessary data across all jobs is available.

The measurement of PSE is readily accomplished when there is only one target job, using either the performance scores themselves (the criterion), or the predictor scores in standard score form multiplied by the validity coefficient. As is described more completely in Zeidner and Johnson (1989a), the mean of the criterion scores and the MPP scores, both expressed in standard score form, are equal in the group that has been accepted (selected).

The equivalent relationship is also true for the more complex forms of personnel utilization; the mean of the actual performance scores is equal to the mean of the MPP scores for each affected group (those selected and assigned to each job or job level), after the selection/assignment process has been completed.

In contrast to traditional (simple, unidimensional) selection, performance scores for those selected and/or assigned to a job cannot, for most personnel utilization processes, be determined by the use of a simple analytical formula. For example, in simple selection, MPP equals the validity coefficient multiplied by the quotient of the ordinate of the normal curve divided by the percent selected, while the comparable value of MPP for a typical classification example is based on a formula involving multiple integrals that are essentially unsolvable. Thus, the increase in efficiency, expressed as a gain in the MPP standard score resulting from the selection/assignment process, is based on the use of more complex personnel utilization processes which take advantage of a multidimensional joint predictor-criterion space. But the increase can only be determined at the price of not being able to use the simple analytical method of computing MPP scores that result from the selection/assignment process.

The computational procedures for operational selection and classification efficiency have much in common. MPP standard scores are equal to the mean of the actual performance scores (expressed in an appropriate metric) multiplied by the validity coefficient pertaining to each job. There are, however, additional computational complications that make classification different from selection in estimating efficiency. For example, while performance scores of those assigned to a set of jobs as the result of a classification process, expressed in standard scores based on the total applicant or assignable population, are adequate for the computation of operational classification efficiency, this procedure does not provide adequate information for computing potential classification efficiency, since scores on all jobs for each individual are required. But invariably only predicted performance scores are available for the computation of potential process efficiency. That is, the option of computing MPP as the product of the validity coefficient and the empirically obtained mean performance scores is not available for the computation of potential process efficiency.

Fortunately, least square estimates (LSEs) of predicted performance scores can be substituted for actual performance scores in both implementing and measuring the effects of the selection/assignment processes. Performance scores are almost never available for an individual across all jobs considered in a multidimensional personnel utilization process.

However, it can be shown that for the covariances among performance scores of multiple jobs, the criterion components that are orthogonal to the joint predictor-criterion space (the space spanned by the performance measures but not by the predicted performance measures) are totally irrelevant to either the implementation of a selection/classification process, or to the measurement of process efficiency. All of Brogden's and Horst's contributions to the measurement or improvement of classification efficiency are dependent upon this fortuitous finding. Both authors independently recognized and extensively utilized this finding, long before a rigorous proof was provided by Brogden (1955). Today no one challenges the substitutability of predicted performance for actual performance.

Thus the least squares estimates of a set of criterion variables (i.e., predicted performance scores) can be substituted for the actual criterion scores employing the correlation coefficients between the predictor variables and the criterion variables. This holds whether the predictors are considered separately or used in a weighted composite. The correlation of predicted performance with actual performance is unity when computed in the joint predictor-criterion space. Most importantly, least square regression weights for the predictor variables remain the same whether predicted performance scores or actual performance scores are used as the dependent variables. Consequently, the same tests would be selected from a pool of experimental tests in maximizing the prediction of either predicted performance scores or actual performance scores. Also, applicants rank ordered on predicted performance would remain in the same order as if they were rank ordered on criterion scores, a consideration that is particularly important in both unidimensional and multidimensional selection.

It is less evident, but equally true, that for personnel assigned to jobs using a linear program (LP) algorithm to maximize MPP scores in the assigned group, the sum of actual performance scores (in standard score form) will be equal to predicted performance scores for those assigned to each job. This follows from Brogden's proof for a similarly stated theorem. Brogden's slightly more general theory states that "for any given assignment of men to jobs, the allocation sum obtained when regression estimates of the criterion are used becomes, as N approaches infinity, identical with that obtained when the criterion scores themselves are used" (Brogden, 1955, p. 252). The term "used" refers to the variables on which the allocation sum (i.e., the objective function) is computed.

Brogden (1955) also showed conclusively by means of a simple algebraic derivation that least square estimates (LSEs) of performance (equivalent to our predicted performance measure) based on all the tests in a battery are optimal (i.e., maximize potential process efficiency) for classification. It is well known that these LSEs also maximize selection efficiency. Brogden's proof is based on the assumption that N, the number being assigned, approaches infinity, and on the further assumption that the best weighted test composites, used both in the assignment process and in computing the objective function, include the total set of predictors that have non-zero regression weights.

An obvious inference can be drawn from the above concerning the Army's aptitude area composites. Each composite consists of three unit-weighted tests (by no means least square estimates) selected from a ten test operational battery. The operational battery, in turn, has been selected from a larger pool of "experimental" tests. Brogden's (1955) proof does not apply to this situation other than to say that these aptitude areas are not the best composites obtainable from the battery or the experimental test pool. There is certainly no evidence that these composites would be equally effective for selection and classification. Brogden's proof, however, underlies almost every classification concept discussed in this volume, including our definition of potential process efficiency that follows.

Abbe (1968) conducted a model sampling experiment to determine the robustness of Brogden's 1955 proof when relatively small values of N are used. The computer generated 10,000 entities (score vectors representing an individual) for two separate analyses divided into 100 groups of 100 entities and also into 10 groups of 1,000, before making optimal assignments using an LP algorithm. Two measures of the objective function, one based on the least square estimate of predicted performance and the other on actual "generated" performance values, did not differ to a statistically significant degree. The results were consistent with Brogden's theoretical proof for infinitely large samples showing that the two measures would provide equal objective function values. These results suggest that Brogden's proof is quite robust with respect to his assumption of an infinitely large N.

Harris (1967) provides strong evidence that Brogden's findings do not apply when "best estimates" are based on only part of the available predictors. As suggested earlier, the reduction in the number of tests in an operational test battery, and the further reduction of the number of tests in a test composite corresponding to a job, or job family, creates a distinction between what is best for use in selection as compared to what is best for use in classification. Such a reduction is almost inevitable in the research and development phase and one must not assume that classification efficiency can be served adequately by the selection of tests for a battery and the use of subsets for test composites designed to maximize selection efficiency.

Performance estimates can be transformed readily to benefit measures by first converting the scale into a metric that is consistent with the cost metric and then providing weights that reflect the values placed on performance in different jobs and/or in the different levels within jobs. Such weights may be based on policy judgments or on evidence bearing on economic value.

Since personnel utilization efficiency is primarily of psychometric rather than economic interest, measures of classification efficiency are usually expressed by psychometricians in terms of performance instead of benefit. The metric conversion and further transformation of performance is deferred until after a number of serious psychometric issues have been examined. As noted, process efficiency is measured in terms of predicted performance. The MPP standard score for selected and/or assigned personnel resulting from a specified personnel utilization process constitutes our measure of process efficiency. Techniques for improving personnel utilization are evaluated in terms of their effect on personnel process efficiency. The best test battery and the best set of test composites are defined as those yielding the highest potential process efficiency.

Potential selection efficiency (PSE) for traditional unidimensional selection may be quite simply measured, as described above, using a function of the validity coefficient, the ordinate of the normal curve at the cut point, and the percentage of applicants who are selected. It is not necessary to compute the MPP standard score as a direct function of the mean criterion score. In contrast, the measurement of operational selection efficiency requires the computation of the MPP standard score for those in the accepted group. The predicted performance scores are standardized to have a mean of zero and a standard deviation of one in the applicant population. If, and only if, the operational selection process differs from the ideal process of rank ordering all applicants on predicted performance and then rejecting all applicants that fall below a given cut score, will the potential and operational measures of selection efficiency differ.

The MPP standard scores of the rejected and accepted groups are related by the requirement that their weighted sum equals zero, the weights being the percent of the applicant population in each group. Thus the MPP standard score used as the measure of process efficiency can be obtained from either the accepted or rejected group, or the separate estimates obtained from each group aggregated into a single estimate.

It should be noted that the performance standard scores used for the computation of PSE and PUE are standardized on the applicant or youth population in the military context. Thus both these indices reflect the gain in the MPP score in the selected, or selected and assigned groups over the MPP score in either the applicant or youth population. In general, the comparison will be made with the youth population when the total process, including recruiting, is being evaluated; the applicant population will be used when it is desired to evaluate the later processes independently of both recruiting policies and procedures and societal effects that determine who of all those in the youth population will become applicants.

Since PCE should reflect the improvement the assignment process can accomplish (making optimal use of the classification test battery) above and beyond selection effects, PCE is defined as the MPP standard score for the assigned personnel; the performance standard scores in the group to be assigned have a mean of zero and a standard deviation of less than one (i.e., the result of truncation or restriction effects introduced by the selection effects on an application population that in many models is assumed to have a standard deviation of one).[2]

It is easy to separate the effects of selection and classification in a two stage process in which selection is accomplished by a single test composite (e.g., a general mental ability measure such as AFQT) in the first stage. In such a process, PSE and PCE are additive (i.e., PSE + PCE = PUE). This is so because our definition of PCE calls for using the MPP score resulting from the selection process as the mean of the performance standard scores that, when averaged after assignment, provides the MPP score used as the measure of PCE. The result of the selection process is the starting point of the classification process, and the result of the classification process is the combined result of both selection and classification. The values for PSE, PCE and PUE reflect this sequence.

When selection and classification are to be simultaneously and optimally accomplished as a single integrated process, the separate consideration of PSE and PCE is not meaningful; one can only measure the results of the integrated process, that is, PUE. The effects of selection can be examined only by computing PUE separately for various selection ratios.

---

2 See Appendix 1B for an example in which an application population is corrected for the effect of a truncation of the left tail.

We have referred previously to an algorithm for the simultaneous and optimal selection and classification process as the multidimensional screening model (MDS). A similar, almost equivalent, process to the MDS can be described as follows: (1) make a trial assignment, by means of an LP algorithm, of the entire applicant population, using quotas proportional to the desired number to be selected and assigned to each job; (2) rank order all applicants on the predicted performance measure corresponding to the job to which the individual has been tentatively assigned; and (3) identify a cut score on each predicted performance score continuum such that the desired quotas will be met by accepting everyone on that continuum who has an equal or higher score.

There is obviously no justification for using such a multidimensional selection and assignment process for selection if the applicant is not going to be assigned to the job for which he or she was tentatively selected. Coupling the selection aspect of an MDS process with random assignment to jobs of the accepted personnel does not provide a useful means of separately estimating PSE, since the predicted performance variables corresponding to the job to which each person is tentatively assigned to effect selection cannot be appropriately used in the estimation of PSE. Instead, the predicted performance variables corresponding to the job on which each individual is actually assigned would have to be used to compute the MPP scores to be used as an estimate of PSE. In such a random process, PCE would be zero and PSE would not be high enough to make the MDS process attractive as a selection process. Thus assuming random assignment in the computation of a PSE using an MDS-like process is not recommended.

Classification efficiency includes either or both of the hierarchical layering and allocation classification effects that may be present and utilized by the classification process. To capitalize on and measure hierarchical classification effects, a multiplicative weighting of predicted performance scores may be used to reflect importance or value accorded to each job or job level. Predicted performance scores are standardized and may be multiplied by their validities prior to the application of these weights. The overall MPP score reflecting the process efficiency measure is then obtained by averaging the MPP weighted scores (if value weights are used) of those assigned to each job, weighted by the number assigned to each job. In determining PCE, the maximum available information is used for computing predicted performance scores contributing to the MPP score for the final outcome (the PCE value). The assignments contributing to this determination must be made using an optimal assignment process. To be fully optimal this process must use as the assignment variables the set of least square estimates that make full use of the test battery or pool of experimental tests for which potential efficiency is being measured.

PAE can be measured in the same manner as PCE if HC effects are removed; that is, if only allocation effects are present in the classification process. One way to assure that the classification effects are due only to allocation effects is to ensure that the least square estimates used to make assignments have equal means and variances in the population being assigned, and remain unweighted with respect to either job validities or values. PAE will be zero if the dimensionality of the joint predictor-criterion space is one, since the prescription of equal means and variances for all test composites used in the assignment process prevents the assignment process from capitalizing on any inequalities of MPP scores across jobs that are due to differences in either validity or value weights. If the classification process capitalizes on "hierarchical layering effects" present in the data, PCE will exceed PAE. The difference between PCE and PAE might appear to be an appropriate measure of hierarchical classification efficiency; unfortunately, hierarchical classification and allocation effects are competitive when both are present and it is clear that HC effects and allocation effects by themselves may approach the contribution to PCE provided by the presence of both effects. The interaction of HC and allocation effects is additive to only a very small extent in the achieving of the total PCE. This is illustrated with our four variable model in Appendix 1B.

## E. ASSIGNMENT APPROACHES FOR SELECTION AND CLASSIFICATION

The personnel utilization process can take place in a single integrated stage or in two or more stages in which the last stage(s) are classification and/or placement processes. The military services have traditionally separated the process of personnel utilization into two stages: a first selection stage and a second classification stage. This is done, in part, because Congress mandates use of a single selection instrument (the AFQT) to determine eligibility to enter the service. During the period of the draft, the AFQT both metered and distributed manpower quality across the services.

In some military training programs evaluation takes place in such a way as to constitute a multiple hurdle process (e.g., the Army helicopter pilot training program). The use of separate criterion components, along with varying costs associated with administering separate types of predictors, can also lead to a multiple hurdle selection process. The use of a multiple hurdle process reduces selection efficiency, as compared to the use of a single least square estimate, and also complicates (but does not prevent) the determination of operational and potential selection efficiency. The computation of a PSE index, when a multiple hurdle selection process is used, requires a correction for restriction in range after each hurdle has taken its toll of the applicants. This correction must be accomplished before computing the MPP score used as a measure of PSE at each selection stage. If predicted performance scores are standardized to have a mean of zero at each successive selection stage, the PSE indices are additive across stages; the sum of the PSE indices at each stage will provide a PSE value for the total selection process.

Just as the multiple hurdle approach is sometimes substituted for the regression equation selection process, many selection/assignment processes substitute less efficient algorithms for the maximally efficient ones in order to accommodate personnel policies. In many real situations, policy considerations take precedence over maximization of benefits. Nevertheless, it is highly desirable to identify the processes and predictor sets which provide the greatest potential efficiency and to use the process yielding the greatest potential efficiency as the starting point. Modifications then can be incorporated into the process allowing implementation of policy and consideration of administrative feasibility. In this section, the effect of common assignment algorithms on the attainment of process efficiency is explored.

The most efficient process is one which utilizes the least square estimates of predicted performance, each based on the full battery, as the selection/assignment variable associated with each job, and that uses an algorithm that minimizes the MPP score in the rejected group (when selection is involved), and maximizes the MPP score in the selected and assigned group (when there is more than one job). The need to fill quotas for each job forces a compromise; everyone cannot be assigned to the job in which he or she could do best. However, the requirement can be imposed on a multidimensional selection algorithm that no rejected individual can have a higher predicted performance score for a given job than anyone selected and assigned to that job.[3] A process by which individuals are being selected simultaneously for several jobs can not be considered optimal unless it achieves this objective.

The use of a hierarchical placement process may lead to an increase in MPP, and thus to an improved PUE, and the consequent increase in utility. It has an interesting

---

[3] This condition is not met by any selection algorithm known by us to be in operational use in a multiple job situation; the MDS algorithm described in this chapter does meet this condition.

similarity to the corresponding contribution of a hierarchical classification process. In a placement process in which a single predictor is used to place a set of individuals in hierarchical levels in one job, with the objective of maximizing MPP while meeting quotas, the MPP standard score can be maximized by rank ordering the eligible individuals on predicted performance and selecting from the top down on this continuum until the quota is met, for each level. Rank ordering may be based on validity or other measures of mean predicted benefit for that level.

The following multi-job hierarchical classification illustration closely resembles hierarchical placement with respect to the manner in which the efficiency of the process is computed. In our hypothetical example the PAE is zero (i.e., the joint predictor-criterion space has a dimensionality of one); however, each of seven assumed jobs has its own associated test composite for use in making assignments. The test composites are least square estimates with each composite's standard deviation proportional to its validity. In actual examples of this unidimensional type, the composite weights or test composition may vary somewhat due to error variance interacting with a dimensionality only slightly greater than one. In our example (unlike the Army's aptitude area composites which have equal means and standard deviations) the test composites have diverse means and standard deviations that are proportional to their validities (since they are PP variables) and thus can take advantage of hierarchical classification effects. The validities given the seven jobs are as follows: 0.65, 0.60, 0.55, 0.50, 0.45, 0.40, 0.35. Thirty percent of the applicants are rejected and ten percent are to be assigned to each job in such a way as to maximize the MPP score for those selected, while exactly meeting the quotas. This could be accomplished with an LP program or by a much simpler process described below.

Our simple assignment process, one that is as optimal as an LP program for this example, calls for placing each individual in rank order on his or her predicted performance score corresponding to the most valid or (in other possible examples) the most valued job. The ten percent of the applicants having the highest scores on the composite associated with this job are assigned to this job. The next most valid job is then assigned the highest ten percent of the remaining applicants on the corresponding composite score continuum, and the same process is repeated for each job in order of its validity. Using entries from a normal curve, one can compute the MPP scores of those assigned at each hierarchical layer and thus compute the average MPP score used as an index of PUE. The resulting value for the MPP standard score (PUE) in our example is 0.315 (106.3 in terms of Army standard scores).[4]

One can easily compute the PSE value for the above example under the assumptions that all selected applicants are randomly assigned, and the means and standard deviations are equal across jobs. Using these two assumptions, the validity achievable with a selection process would be 0.50 and the MPP standard score of the accepted group is 0.248. Thus, the gain achieved by hierarchical classification over the use of simple selection and random assignment conditions, where the PAE is zero but validity differences fairly large, is 27 percent.

For multidimensional selection, placement or hierarchical classification, as well as for allocation, either a primal or dual LP program (or an approximating algorithm) is essential to the practical implementation of the assignment process. The designation of the terms primal and dual to a particular linear program algorithm is somewhat arbitrary since the dual of a dual algorithm is the primal algorithm.

It is traditional to designate the simplex and the related algorithms that maximize mean predicted performance (MPP) while meeting quota constraints as the primal version. The simplex algorithm starts with a feasible (i.e., meets all the constraints) but less than optimal solution. This initial solution is referred to as a basis and is the first step of a series of iterative, feasible, solutions (each one more optimal than its predecessor) that continue until the objective function (MPP score) is maximized.

The dual solution corresponding to our primal example would consist of an algorithm which seeks to minimize, as the objective function, the difference between the obtained and desired quotas (the constraint of the primal solution), while constraining each iteration to yield the maximum possible MPP score (the constraint of the dual solution). The dual solution is thus a sequence of iterative solutions in which the MPP score remains a maximum for the set of quotas that are met, but the desired quotas are not met until the last and final solution. The various algorithmic versions of the Brogden-Dwyer optimal regions algorithm (Brogden, 1946b, 1954a, 1954b; Dwyer, 1954, 1957; Boldt and Johnson, 1963; and Larkin, 1966) are the best examples of useful dual LP versions. The dual is especially useful when the approximate meeting of quotas is permissible, scarce resources prevent all quotas from being filled, or selection and assignment is to be

---

4 See Appendix 1B for the detailed computation of this example.

accomplished simultaneously. The major disadvantage of the optimal regions solution is that, unlike the simplex and most other primal versions, the solution is not obtained in a finite number of steps in which each iteration is necessarily better than the last.

Several LP algorithms that alternate between primal and dual solutions in successive iterations are also available. Granda and McMullins (1972) investigated an algorithm which, if the gap between the highest and next highest score exceeded a specified amount (one iteration of a simple dual algorithm), assigned individuals to the job corresponding to their highest composite score. The much smaller remaining group of individuals would then be assigned by means of a more time-consuming primal solution. The group to be assigned with the primal algorithm could be kept quite small to the point where little more than tie breaking was being accomplished, if a small degree of approximation with respect to the objective function was considered permissible.

The optimal regions algorithm has the advantage of its logic being easily understood. The use of this algorithm hinges on the following important theorem: when the correct constant for each job, usually called a column constant, is added to all individuals' test composite scores yielding adjusted scores corresponding to a particular job or job family, the desired optimal solution is obtained by assigning each individual to his or her highest adjusted score (Brogden, 1954a). In short, the use of the correct set of column constants will achieve the optimal solution. Trial assignments to determine how far the quotas have been missed, and the re-estimation of the column constants constitute the successive iterations of an optimal regions algorithm.

The optimal regions algorithm provides a direct and easily understood way to accomplish an optimal simultaneous selection and classification process. The three steps for accomplishing such a process, using a primal algorithm, were described earlier. The most direct way to accomplish a simultaneous selection/classification process would be to modify the Brogden-Weaver algorithm slightly (Larkin, 1966). The modification involves directly seeking the column constants that will provide for the assignment of the correct number to each job. A particular advantage of this algorithm is that a set of quotas that adds to less than the total number of applicants poses no difficulty and hence there is no need for the creation of a rejection category.

The required column constants can be obtained as a by-product of many primal LP algorithms. Thus, some analysts may prefer to use primal LP off-the-shelf software to compute the column constants needed for the selection/assignment process. These column constants can then be closely approximated from inspection of adjusted predicted performance scores (or test composite scores) for which assignments have been designated using a primal LP algorithm. The required column constants are computed by first subtracting each person's largest score from each of his other scores. A person's largest adjusted score will then be equal to zero and all other adjusted scores will have a negative sign. The adjusted scores corresponding to the job to which each person was assigned are rank-ordered, and the adjusted score located at the cut point which will provide the correct number in that job is identified. Each such negative adjusted score is the appropriate column constant to be subtracted (or changed to a positive number and added) to obtain the optimal regions solution. This column constant obtained from the simple computations described above that started with a primal LP solution to the classification problem in an applicant sample may approximate the exact quotas desired for both selection and assignment closely enough for operational purposes, when the same column constant is used both to select and to assign; if not, further exactness can be achieved in one or more further iterations using one of the Brogden-Dwyer algorithm versions.

Prescribed personnel policies may prevent the use of off-the-shelf LP algorithms, or the use of existing computer programs, in implementing a classification process. This may be the case, for example when: (1) there may not be a sufficient number of individuals in the assignment pool to meet all the quotas; (2) policies may require two or more objective functions to be successively maximized using the slack left over from the prior optimizations to achieve the later ones; (3) constraints may need to be prioritized by policy when all constraints cannot be met; or (4) constraints may need to be successively relaxed, in accordance with priorities prescribed by policy, until a feasible solution can be obtained. In general, such complications are dealt with by modifying LP algorithms to such extents that modified programs commonly assume a name of their own.

One such class of algorithms, goal programming, accomplishes a constrained optimal solution by establishing a hierarchy of objective functions (e.g., travel cost, meeting applicant preferences, accomplishing a desired distribution of quality into the various job families, the MPP score), optimizing objective functions in the indicated order, with a high probability that all the slack required for further progress will be used before the end of the list is reached. Considering that most of these complications will have the effect of reducing the MPP standard score, the difference between operational and potential classification efficiency is enlarged through their use; competing constraints and objective functions can only reduce MPP.

The search for simplicity in assignment algorithms also increases the gap between operational and potential classification efficiency. For example, during the early years of the all-volunteer force the Army had great difficulty meeting recruiting quotas; consequently the then existing complex LP driven assignment system was discontinued. The use of cutting scores on aptitude areas once again was prescribed for use as the only source of operational classification efficiency. Assignments largely were determined by what the recruiter could sell to a potential recruit. The recruit, upon acceptance, entered the Army with a contract that specified either the job family or the geographic area to which he or she would be assigned. The recruit needed to meet only the minimum aptitude area standard for a job.

## F. THE INITIAL IMPLEMENTATION AND EVOLUTION OF A CLASSIFICATION PROCESS

As mentioned in Zeidner and Johnson (1989a), the introduction of the Army Classification Battery in 1949 was a major innovation for military personnel utilization. The ACB was developed with differential classification in mind, to capitalize on inter-and-intra individual differences. The origin of the military's current use of a two-stage selection and classification process originated just before the introduction of the ACB. The history of the use of the ACB, and later the ASVAB, is a case study of personnel policy impact on test battery usage.

From the time of implementation of the ACB, the operational classification process greatly underutilized its classification potential. In the mid 1960s computer and software technology developed to the point that the use of an LP algorithm for large-scale assignment became practical and soon after an LP capability was installed. But the perceived need to reduce costs, meet job preferences, and distribute quality appropriately, left little room to maximize MPP. It should be noted, however, it was the use of differential selection and separate cutting scores for each job, that the developers of the ACB visualized as the enabling mechanism for the classification process. It was this mechanism that was relied upon to make the ACB more efficient than its predecessor, the Army General Classification Test (AGCT), a single selection and placement test.

Just before the change to the ACB, the AGCT was used in two stages, first for selection, and then for hierarchical classification to jobs. The classification process used cutting scores corresponding to a school course or an MOS (course/job) hierarchy; the probability of failure was minimized by using higher cutting scores for course/jobs having higher validities and/or failure rates. The technical school courses had higher validities and failure rates than the combat arms courses. Consequently the average minimum required cutting score for the former was much higher; considerable dissatisfaction was expressed however, because the combat arms were not receiving adequate numbers of high quality personnel.

The minimum required aptitude area scores for Army school courses and some MOS were at one time determined for operational use on the basis of research data. These cutting scores were defined as the point where fifty percent of the soldiers were predicted to be unsatisfactory (i.e., one half of all soldiers with that score could be expected to fail the preparatory school course). Cutting scores computed in such a way reflected both the magnitude of the validities and the difficulty of the course. In earlier years, cutting scores tended to be considerably higher for technical MOS than they are today. They were drastically lowered in the early post-Vietnam era of the all-volunteer force when the Army experienced shortages in higher quality personnel. At one time cutting scores were further reduced to ease the entry of minorities into the Army and to reflect the prevailing view that all recruits were trainable, "trainees did not fail; the trainers failed," and none but outright disciplinary cases should fail. Even at present it is clear that cutting scores are much lower than they should be if the tests are to provide adequate classification effects in the absence of an LP-type assignment program. It is indeed fortunate that the new Enlisted Personnel Allocation System (EPAS), mentioned in Zeidner and Johnson (1989a) and more fully described in Zeidner and Johnson (1989b) is now under development. EPAS will no longer rely so completely on cutting scores.

The use of cutting scores on a single instrument, such as the AGCT, could not provide above average MPP scores to some groups without assuring that other groups would have balancing below average MPP scores. The possibility of capitalizing on intra-individual differences in predicted performance scores, assigning individuals according to their higher scores as often as quotas permit, offered an attractive solution to this problem. As noted earlier, using ACB aptitudes area composites, as many as 80 percent of the recruits could be assigned to jobs where their predicted performance scores would exceed the average performance in a randomly assigned population. This potential, if realizable, could have solved the quality distribution problem that was plaguing the combat arms.

ACB based assignments were accomplished initially by military counselors at each basic training camp that met quotas for the assignment of soldiers to school courses and other training after basic training. It was intended that each soldier would be assigned to a job family corresponding to one of his two highest (of ten) aptitude area scores, but no special effort was to be made to achieve an assignment to the higher of the two. Each school course or on-the-job trained MOS had its own aptitude area cutting score that also had to be met. Consideration by the counselors of these two factors provided some gain in the operational classification effectiveness, but this gain fell far short of the PAE of the battery.

Sometimes counselors, in making classification decisions, give overriding consideration to factors other than predicted performance. These include: (1) the reduction of travel costs from basic training to the next assignment; (2) the matching of soldiers' preferences; (3) the meeting of a variety of job or course prerequisites (e.g., prior completion of a course in trigonometry, required physical profiles, complete absence of adverse court records or of color blindness, and a required length of remaining enlistment); (4) the distribution of quality (e.g., increasing the number of persons with higher aptitude area scores assigned to the combat arms). These considerations, singly or combined, inevitably reduce the MPP score resulting from classification.

The decentralized person-job matching system in which the counselor made the final decision in the presence of the basic trainee was later centralized to a Pentagon location and mechanized to the extent of placing each individual's information on a Hollerith card. Sorters were utilized in a cascading approach to identify assignments. At least one additional constraint was introduced: the combat arms were given the same proportion of college graduates as the other Army branches in initial assignment.

The developers of the ACB wanted to achieve more of the PAE inherent in the battery through use of the battery in conjunction with the Brogden-Dwyer optimal regions algorithm. The algorithm was first described by Brogden in 1946, and then presented as a more precisely described algorithm by Dwyer in 1953. It was not surprising that Brogden, in the early 1960s, encouraged research to program and demonstrate an LP type assignment process.

An improved version of the Brogden-Dwyer algorithm, the Brogden-Weaver algorithm (Boldt and Johnson, 1963) was devised and programmed on the IBM 1401 computer, a relatively small computer primarily used for I/O support of the larger IBM 705 personnel data processing computer. Using the Brogden/Weaver version, a near-optimal solution for 3,000 soldiers and 75 jobs was determined in a little more than two hours. All the constraints required of the sorter operation were implemented except that quotas for some jobs, designated by policy, were defined by a range, a compromise that also sometimes occurred in the sorter supported process. The improvement in the MPP was undoubtedly considerable, although not documented to the extent desirable (Boldt and Johnson, 1963). It was unfortunate that results for the operational assignments comparable to those available for the demonstration were not provided by the operational office conducting the sorter supported process; the results from the demonstration could only be compared to less comparable operational data as described below.

A comparison was made between the aptitude area mean scores resulting from a demonstration of the optimal regions method using 5,128 enlisted men (primarily draftees) in January, 1961, and those resulting from the sorter supported process using 1,204 draftees entering during October of 1959 (Boldt and Johnson, 1963). The comparison was made in terms of Army aptitude area composites standardized to have means of 100 and standard derivations of 20 for the tests comprising these composites in a youth population; the aptitude areas had standard derivations ranging from 17 to 21 in the recruit population. The aptitude area mean scores ranged for the computer assigned group from 109 for "infantry" to 121 for "general technical." The sorter supported assignment yielded a range from 98 for "infantry" to 114 for "clerical," and only 103 for "general technical." Except for "clerical" which had the same result for both groups, the gain for the computer assigned group was not less than 8 points (for "electronics") and averaged 11 points. Even after noting that these two groups obtained at different times of the year would not have been compared if any other data could have been obtained, one is struck with the potential this approach had for both solving the quality problem in the combat arms and increasing the MPP score for first job assignment. Also there is little doubt that the number of human errors in the assignment process would have been significantly reduced.

There was a growing recognition in the mid 1960s that a computerized optimal assignment model was desirable. The Marine Corps became highly interested in the Army results and developed a quite different primal LP algorithm, one that was both more flexible and efficient for implementing their objectives. This procedure was successfully used to make operational assignments in the Marine Corps (Hatch, 1966, 1970). Encouraged by the success of the Marine Corps, the Army utilized an assignment computer program that evolved into a full-blown system, the "ACT II," which later provided the basis for both the Air Force and Navy classification systems, long after the Army ceased using ACT II for making initial assignments.

ACT II was both an efficient and flexible classification process that had many of the features of goal programming. It permitted the sequential optimization of successive objective functions (e.g., transportation costs, MOS preference, aptitude area scores), the sharing of quota shortages, and the sequential relaxation of constraints. For example, all constraints could be successively relaxed to permit the assignment of the entire personnel pool. An impressive degree of flexibility existed for implementing policy changes without reprogramming the system (Hatch,1970).

ACT II appeared to be heading toward a happy future, considering the capabilities available from its software. Unfortunately, policymakers were not confident that the Army should sacrifice the savings of travel costs for increased MPP nor to deny an enlistee (as contrasted to a draftee) his job preference if he met the required minimum score. As the changeover to an all-volunteer Army occurred in 1973, the sophisticated ACT II features for accomplishing centralized batch assignments were no longer useful. What was needed instead was up-to-date information on which quotas were open and a means for reserving a slot for a specific MOS immediately upon extending a promise to a new recruit; an information and communication system evolved rather than a decision system. Thus, EPAS was initiated to fill a vacuum rather than to improve an LP driven classification system already in use.

It would appear that the era of the batch primal LP program came to an end, gloriously enough, with the demise of ACT II. Recruiting needs precluded the luxury of batch assignments. Fortunately, however, computer and communications technology has now advanced to the point where person-by-person dual LP programs can be part of a combined, simultaneous recruiting and assignment system. The required concepts have been available since 1946 (Brogden, 1946b).

## G. DECENTRALIZED CLASSIFICATION: PERSON-BY-PERSON ASSIGNMENT ALGORITHMS

Making assignment decisions for a recruit or soldier without waiting to accumulate a large enough personnel pool to justify use of a batch algorithm is referred to as a person-by-person algorithm. It is also called sequential assignment by the Air Force, and line-by-line assignment by others. Such an algorithm can provide an exact solution for the defined population as N approaches infinity; the quotas for this defined population will usually be estimated as the desired input that mirrors requirements, modified by insights into the economy and demography of the nation that may force a compromise between requirements and recruiting estimates. The quotas contained in a batch program, such as ACT II, were similar estimates of future requirements.

A practical person-by-person assignment process, using Brogden's concept of additive column constants, could be provided by computing, on a weekly basis, the next four weeks of estimated input, as either a pool of synthetic, generated entities, or as a covariance matrix which could in turn be used to generate synthetic entities. (See Chapter 4 for a description of the procedure.) This pool of entities could then be used as the data to obtain an LP solution meeting specified quotas and other constraints. The resulting column (job) constants would then be used for one week to make person-by-person assignments. A job constant would be added to the corresponding test composite and these adjusted scores for each job compared within the individual. Assignment to the largest adjusted score is an optimal assignment for the defined population and the recruit. In the event that the recruit cannot be given an optimal assignment, the recruiting or assignment counselor can readily see the penalty incurred on the objective function exacted by each alternative choice of jobs.

In a person-by-person assignment process, the quotas for the start of a particular course would not automatically be met. There would be an obvious need to adjust the reporting dates of the recruits to meet quotas on specific start dates for courses.

The use of a batch LP program, whenever a specified number of applicants is accumulated, would not only delay the decision process, and possibly result in the loss of some potential recruits, but also could be expected to make poorer decisions with respect to the input population, reflected by a lower objective function value than provided by the above person-by-person algorithm. The above statement presumes that the "batch" assignments are optimized with respect to fluctuating constraints (e.g., quotas, quality goals, etc.), and weekly or biweekly input characteristics. In contrast, the person-by-person algorithm is based on the population constraints and input, but quotas would necessarily be imperfectly met over small time periods although closely approximated over the sum of these periods. Thus we see that the practicality of a person-by-person assignment algorithm depends on being able to make some assignments from a waiting list in order to meet weekly enrollment goals for individual school and training courses.

Horst (1960) and Sorenson (1965b) proposed a person-by-person assignment process that would use a multiplier matrix converting each applicant vector of test composite scores into a surrogate assignment vector approximating one row of the assignment matrix. In each row of such an assignment matrix, one element is unity and all others are zero. The applicant would then be assigned to the job family corresponding to the highest element in the surrogate assignment vector. This method has the advantage that the required transformation matrix can be computed directly from the matrix of covariances among the test composites in the input population and the validities, without the generation of a pool of entities and the conduct of a simulation. For the previous method [the use of column constants applied on a line-by-line implementation of Brogden's (1946b, 1954a, 1954b) algorithm] the difference between the highest adjusted score and the adjusted score corresponding to the alternative, less optimal job to which an individual was assigned had meaningful implications regarding the resulting reduction in the objective function value. The significance of such a difference between the best and an alternative assignment in the Horst-Sorenson method is not known.

Ward (1958) proposed the use of a disposition index (DI) that could be used by counselors required to make assignments. This computation of a set of indices to be used with an individual being counseled, one DI for each job being considered, would require knowledge of the set of predicted performance scores for the individual at hand, the MPP score (across jobs) for the individual, the MPP scores for each jcb (across individuals), the overall MPP score for the expected input over some prescribed time frame, and both the number of individuals and the number of jobs to be considered in the designated time frame. Assignment of the individual to the job corresponding to his highest DI was recommended.

In his two introductory examples, Ward used three individuals and three jobs, with the quota for each job being one. His final example had the objective of minimizing cost, rather than maximizing performance, and consisted of three categories of people (of unequal numbers) to be assigned to five jobs with unequal quotas. In this last example, personnel were to be assigned to a job in ascending order of their lowest DI. This proposed algorithm required a batch mode for its implementation; otherwise there was no provision for meeting quotas. Alternatively, frequent recomputation of the DIs would be required. This algorithm appears to have no obvious advantages and some apparent disadvantages compared to the direct addition of column constants to job performance estimates and the assignment of each individual to his highest adjusted score. The latter procedure produces a maximum MPP score and can be made to produce the exact quotas by appropriate scheduling of recruits into basic training.

By the 1980s, Ward's DI approach had been refined and incorporated into the optimization module of the Air Force personnel acquisition and assignment system (PROMIS/PJM). This "sequential" process provides an assignment for one person at a time, the optimization decision taking place within the context of having only one person and many jobs. Predicted competition is estimated and a modified DI calculated so that the jobs can be rank ordered in terms of highest to lowest total system payoff. The relative importance of each potential person-job match for an individual could also be determined as an optional capability, if desired. The present effectiveness of the Air Force sequential assignment system is due primarily to Ward's contributions.

Cardinet (1959) proposed a graphical person-by-person assignment aid which could be overlaid onto a graphical (profile) display of each individual's predicted performance. Cardinet assumed that a counselor would be more comfortable with the use of profiles; thus a method of combining accurate classification with a non-demanding process was provided to the counselor. The principal disadvantage of this approach was the effort and cost required to develop the standard profiles to represent each job (or job family), and to represent each applicant's set of scores as a profile.

Brogden (1954b) is cited by Cardinet as the source of the concepts that stimulated the development of his approach. As Cardinet pointed out, in comparison with Brogden's method, the advantage of the profile is the ease with which it can be applied by the counselor. The same results would be obtained by adding the appropriate column constants, one corresponding to each job, to each predicted job performance value and recommending those jobs with the higher adjusted scores. Similarly, the counselor could refer to a table containing minimum requirements for each job to determine if the applicant can be selected for the job to which he would be assigned optimally if minimum eligibility for that job is established. Only those applicants lacking eligibility for any open job would be rejected.

Cardinet proposed the use of his standard profiles for multidimensional selection and classification. For the former (called differential selection by Cardinet at one point and multiple selection at another), "a minimum is fixed separately for each predicted success, and a subject is eliminated if he does not reach the minimum in any job" (Cardinet, 1959, p. 197). If a candidate is selected he is then assigned to the job identified by comparison of the individual's profile with the same standard profile used for selection. The individual, in effect, is assigned to the job in which his predicted performance exceeds the minimum score used for selection by the greatest amount. This is entirely consistent with the MDS process described earlier in this chapter.

# APPENDIX 1A
# A SIMPLE MODEL OF HIERARCHICAL LAYERING

## APPENDIX 1A.1: CONCEPTS AND NOTATION FOR THE CHAPTER 1 APPENDICES

Although the other chapter appendices make extensive use of matrix algebra, only techniques and concepts taught in elementary algebra and statistics courses are utilized in the appendices of this chapter.

Approaches for computing MPP resulting from optimal assignment to two jobs in situations in which pure allocation, pure hierarchical classification, or a combination of the two provide the classification effects are described in Appendix 1B. This approach is demonstrated using values for validities and the intercorrelation among the two assignment variables that are reasonable to expect in practice. We hope that these results will sharpen the reader's intuition as to what may obtain when there are several, or even many, jobs and corresponding assignment variables. Techniques for measuring available classification efficiency when assignment is to be made to more than two jobs will be provided in Chapter 4.

Appendix 1A focuses on the situation in which classification to two or more jobs is to be accomplished using a single predictor with disparate validities. Assignment is to be accomplished by use of predicted performance (PP) scores– the product of the predictor score in standard score form and the validity coefficient. The objective of optimal assignment is the maximization of the mean predicted performance (MPP) standard score. The MPP score output by our model is this maximized value of MPP.

The operational situation being modeled in each example will be defined in terms of the validities ($R_j$) of the predictors of the j[th] criterion. When there are two or more predictors, as in Appendix 1B, each assignment variable is a best weighted test composite ($R_j$ is a multiple correlation coefficient) and there is a common correlation coefficient ($r$) among the pairs of PP predictors. In this appendix (1A), the correlation coefficient among PP scores ($r$) is 1.0 since there is only one predictor variable. This constraint on r permits the use of a simplified model to compute MPP for such a hypothetical operational situation involving any number of jobs. When $r$ is less than 1.0 we must use a more complex model and restrict ourselves to the study of hypothetical situations containing only two jobs.

We use the same notation in Appendix 1A and 1B as we use in Chapter 2 appendices to describe Brogden's 1959 model: $r$ represents the correlation among PP scores; and $R$ their validities (when equal across jobs). Validities are also represented in the format $r_{aA}$ where $a$ is the predictor and $A$ is the criterion.

In Appendix 1B we define a more general model in which the predictor need not be a predicted performance measure, but all of our hypothetical examples and computational demonstrations use computing formulae that assume each pair of criterion variables have corresponding LSE variables, $a$ and $b$. Thus $r_{aA}$ and $r_{bB}$ are multiple correlation coefficients that are also the standard errors of $a$ and $b$, respectively. Specifying $a$ and $b$ as LSEs permits us to compute $r_{aB}$ as $r_{ab}$ times $r_{bB}$, and $r_{bB}$ as $r_{ab}$ times $r_{aA}$. Our model needs only the selection ratio (SR) and the triplet ($r_{ab}$, $r_{aA}$, $r_{bB}$) as model input to completely define the hypothetical operational situation. In the more general model of Appendix 1B, in which a and b are not LSEs, the model also requires as input, $r_{aB}$, $r_{Ba}$, and the standard deviations of both $a$ and $b$ ($S_a$ and $S_b$); both $S_A$ and $S_B$ are also required when SR is less than 1.0, but can be readily computed from the previously cited input.

Thus we see that all of our examples presented in both appendices of this chapter require knowledge only of $r$, each $R_j$ and the SR to define the operational situtation and to accomplish the computations required by the model. Assuming a normal distribution of the predictor variables and the use of LSEs as predictors, all other values required by the models (algorithms) for outputting MPP can be computed from these input values.

## APPENDIX 1A.2: HIERARCHICAL CLASSIFICATION MODEL AND EXAMPLES

This appendix describes a simple approach for optimally assigning personnel on the basis of a single variable used for both selection and classification. When this single predictor variable has disparate validities across jobs, and the continuum of predictor scores is matched against hierarchical layers of jobs rank ordered on the magnitude of the predictor validities, hierarchical classification is occurring. The job having the highest validity and a quota of $n_1$ receives those $n_1$ indivuals having the highest predictor scores, the job with the second highest validity and a quota of $n_2$ would receive the $n_2$ unassigned individuals with the second highest test scores. The predictor test continuum is marked off from the top down, dividing the continuum into layers, or score intervals, until all personnel are assigned to a job.

This process of matching hierarchical layers of rank ordered personnel and jobs accomplishes an optimal assignment of people to jobs, a process which maximizes the mean predicted performance while meeting job quotas. The hierarchical layering solution provides the same solution (the same set of personnel assignments) as is provided by using a linear program (LP) to assign personnel to jobs on the basis of predicted performance (PP) scores--maximizing an objective function of mean predicted performance (MPP) standard scores.

For N individuals that were assigned by a hierarchical layering process, assuming a normal distribution of PP scores, we can easily compute the mean PP score for each job (i.e., for each interval or layer of the continuum), using values from a normal curve table. Letting $n_1/N = p_1$, $n_2/N = p_2$, ...., $n_m/N = p_m$, for $m$ jobs, we commence with $p_1$, the upper tail of the normal curve. At the point on the abscissa, $x_1$, that cuts off an area of the normal curve equal to $p_1$, we label the normal curve ordinate corresponding to $x_1$ as $z_1$. The required mean for the top layer of the predictor continuum is $z_1/p_1$, the mean for the second most valid job is $(z_1 - z_2)/p_2$, and the general term for the predictor mean of the j^th^ most valid job is $(z_{j+1} - z_j)/p_j$, with $j$ ranging from 1 to $m$.

The sum of of these interval means, weighted by the validity of the $j^{th}$ job and $p_j$, is equal to the MPP standard score. That is, MPP = $\Sigma_j^m (z_{j+1} - z_j)R_j$. The interval means for intervals lying primarily below the mean, of course, have a negative sign.

The HC classification example described in the text has a selection ratio of 0.7 and seven jobs, each of which has a quota of N/7. The validities of the seven jobs are: 0.65, 0.60, 0.55, 0.50, 0.45, 0.40, 0.35. Interpolating the table entries from a normal curve table provides values for $z_1$ through $z_7$ as follows: 0.17543, 0.27989, 0.34771, 0.38637, 0.3989, – 0.38637, – 0.34771. Using these values in the formula provided in the paragraph next above yields a MPP standard score of 0.315. We now compare this result with the magnitude of the MPP that results from these validities, optimal selection with an SR of .7 and a random assignment of personnel to jobs.

Random assignment of the selected upper 70 percent of the PP continuum can be depicted as a single interval, $[(z_1 - z_7)/0.7]\ R$, where $R$ is the average of the above validities (i.e., 0.5). The mean of this single interval of selected individuals is 0.4967.

Thus, MPP for the random assignment case is 0.4967 times 0.5 or 0.248. HC classification effects can be identified as the difference between 0.315 and 0.248, or 0.067.

Only a two job situation can be evaluated using the general model of Appendix 1B which permits a complete range of values for r. Using the relatively simple situation described above for two job criteria, $A$ and $B$, and the correlation between the two predictors set to 1.0, we apply our more general model and note that we obtain the same results. Our common example to be compared across the two appendices calls for a validity of 0.6 for one job and 0.4 for the other, an SR of 0.7 and the correlation between predictors, $a$ and $b$, of 1.0. The mean of the tail containing the upper 35 percent of the population is 1.05826, while those in the next 35 percent of the continuum have a mean of $-0.0648$. Thus the MPP standard score for this two job example can be computed as follows:

$$\text{MPP} = (0.6\ (1.05826) + [0.4\ (-0.0648)]/2 = 0.3045.$$

Comparing the above MPP value with that obtainable with random assignment provides a gain in MPP attributable to HC of 0.0562. Comparing this result with that obtained for the seven job example suggest there is little or no gain to be expected from adding more jobs to a HC classification situation. This is in contrast to the major increase in classification effects obtainable from an increase in the number of jobs when allocation effects are present. Brogden's 1956 model shows an increase of MPP by a factor of 2.4 for an increase in the number of jobs from 2 to 7, when the SR equals 1.0 and the classification effects are purely allocation. This increase would be by a factor of 1.7 when SR = 0.7.

# APPENDIX 1B
# A FOUR VARIABLE MODEL FOR EVALUATING ALTERNATIVE CLASSIFICATION STRATEGIES

## APPENDIX 1B.1: THE GENERAL MODEL

In this appendix we describe a model for evaluating the utility resulting from using two test composites to optimally assign personnel to one of two jobs. We then use this analytical model to evaluate the effect of several patterns of predictor characteristics on utility as measured by mean predicted performance (MPP).

We commence with a general formulation of our model in terms of two assignment variables, "$a$" to be used as a measure of predicted performance in a job for which the performance criterion is the variable "$A$", and "$b$" a variable which has the same relationship to the second job (a job with the criterion variable "$B$"). Considerable simplification of the general model results from specifying that a is a least square estimate of $A$ based on all predictors, making "$a$" an FLS composite; "$b$" is similarly related to $B$. Further simplification occurs from either making the validities of the two assignment variables equal to each other, or by making the two assignment variables perfectly correlated although with differing validities of $a$ against $A$ and $b$ against $B$. These characteristics are of interest since they define processes of pure allocation and pure hierarchical classification, respectively.

We use our four variable model to demonstrate utility effects of optimal assignment under two separate selection/classification processes: (1) using a selection ratio of 100 percent, assuming the total multivariate Gaussian distributed population is entirely assigned to the two jobs; and (2) using a fifth variable, g, on which to truncate input, with a selection ratio of 70 percent, again assuming a multivariate Gaussian distributed applicant population. In process (2) selection will be made on variable $g$ for both job $A$ and job $B$; this is in accordance with a two-stage process for sequential selection and classification. We show that three examples used to demonstrate process (1) can be directly verified against Brogden's results (1959, p. 189).

Our four variable model is based on the concept that a cutoff score on the continuum, $d = (a - b)$, divides between those persons appropriately assigned to job $A$ or to job $B$. Any desired pair of quotas for the two jobs can be obtained by an appropriate selection of a cutoff score on $d$. We have arbitrarily chosen *a 50 percent split between the two jobs* for use in all of our examples.

The mean predicted performance standard score for those assigned to job $A$ can be denoted as $(MPP)_A$. The mean of *the criterion A* for those assigned to job $A$ can be denoted, for a quota of $q$, as $M_q$; $M_q = z_q/q$ where $z_q$ is the ordinate of the normal curve at the cutoff point on $d$. $M_p$ is similarly defined as $z_p/p$ where $z_p$ is the ordinate of the normal curve at a truncation point on each of the selection variables ($a$ and $b$). The selection ratio is represented by $p$.

$S_d$ is the standard deviation of $d$ and $r_{da}$ is the product moment correlation coefficient between $d$ and $A$. Using this notation, $(MPP)_A = (r_{dA}\ M_q) + r_{aA}\ M_p$ , and $(MPP)_B = (-r_{dB})\ M_{1-q} + r_{bB}\ M_p$. The value of $M_p$ will be zero when SR = 1.0. When SR > 1.0, $M_p$ is non-zero and both $r_{dA}$ and $S_d$ must be corrected for direct selection effects resulting from the truncation on the selection variable. This correction process will be discussed in Appendix 1B.3.

Our four variable model is essentiallly represented by the above computing formula for $(MPP)_A$, the corresponding formula for $(MPP)_B$ and the aggregate of these two values into $(MPP)_t$; $(MPP)_t = q\ (MPP)_A + (1-q)\ (MPP)_B$. The source of this computing formula is a formula for the biserial correlation coefficient. Using our notation, this formula can be written as $r_{dA} = [(MPP)_A - M_p)/S_A\ M_p$; $S_A = 1.0$. In this equation $r_{dA}$ can be either a biserial coefficient, or, if normality assumptions are met, can be a product moment coefficient. Inserting a value for $r_{dA}$ and solving for $(MPP)_A$ provides the basic formula for our model as follows:

$$(MPP)_A = r_{dA}\ M_q\, S_A + r_{aA}\ M_p \quad . \qquad (1)$$

Similarly, we can compute $(MPP)_B$ using a reversed $d$ to compute what is essentially $-r_{dB}$; thus our other basic formula is

$$(MPP)_B = r_{dB}\ M_q\ \ S_A + r_{bA}\ M_p \quad . \qquad (2)$$

We can also reflect the effects of weighting $a$ and $b$ to provide either variances proportional to their validities or equal variances across the two predictor variables. The latter situation assures that all assignment effects are free of a contribution from hierarchical layering, and all classification effects are thus pure allocation. Our demonstration is solved with and without hierarchical layering, i.e., hierarchical classification (HC) effects.

We make several assumptions to provide simplification of the basic formula appropriate for use in several of our examples. First, we assume that $a$ is an FLS composite providing an LSE of $A$, and $b$ is similarly a LSE of $B$. Thus $r_{dA} = r_{da}$, $S_a = r_{aA}$, and $r_{ab} = r_{bA}/S_a = r_{aB}/S_B$. When we set values for $S_a$, $S_b$, and $r_{ab}$ we are also setting values for $r_{bA}$ and $r_{aB}$, values required by the basic formulation of our model. Note that $S_a$ is also the validity of against $A$ ($r_{aA}$), and $S_b$ is the validity of $b$ against $B$ ($r_{bB}$). We arbitrarily set $q$ equal to 0.5 for all examples, although our model could easily be used to examine the effect of unequal quotas on utility.

Defining each of 8 conditions in terms of $S_a$, $S_b$, and $r_{ab}$ we can compute $(MMP)_t$ using the following equations:

$$r_{dA} = (r_{aA}\, S_a - r_{bA}\, S_b)/S_d\, S_A); \tag{3}$$

$$S_A = 1.0;\; r_{bA} = r_{aA}\, r_{bA} \quad .$$

$$r_{dA} = (r_{aA}\, S_a - r_{ab}\, S_a\, S_b)/S_d \quad . \tag{4}$$

$$S_d = (S_a{}^2 + S_b{}^2 - 2\, r_{ab}\, S_a\, S_b)^{1/2} \quad . \tag{5}$$

## APPENDIX 3B.2: DEMONSTRATING MODEL ASSUMING NO SELECTION; THE FOUR VARIABLE MODEL WITH SR = 1.0

In the special case where $r_{aA} = r_{aB} = S_a = S_b = R$, we see that the numerator for $r_{dA}$ is equal to $R^2\,(1 - r_{ab})$ and the denominator, $S_d$, is equal to $(2)^{1/2}\, R\,(1 - r_{ab})^{1/2}$. Simplifying and inserting into the basic equation for our model yields: $(MPPA)a = R\,(1 - r_{ab})^{1/2} M_{50}\,(2)^{1/2}$. The conditions defined by setting $r_{aA} = S_a$ and $r_B = S_b$ corresponds to Brogden's model (1959) when there are only two predictors and all applicants are selected and assigned.

It is useful to know that simplifying the basic formula for $r_{dA}$ by setting $S_a$ and $S_b$ equal to 1.0, while permitting different values for $r_{aA}$ and $r_{bB}$, yields the same simplified model defining equations as results from setting $r_{aA}$, $r_{bB}$, $S_a$, and $S_b$ equal to the average of $r_{aA}$ and $r_{bB}$. In the general equation for $r_{dA}$, with $S_a = S_b = 1.0$, the numerator simplifies to $r_{aA}\,(1 - r_{aB})$ while the denominator, $S_d$, simplifies to $(2)^{1/2}\,(1 - r_{ab})^{1/2}$. We note that $r_{dA}$ is equal to $r_{aA}$ times $[(1 - r_{ab})/2]^{1/2}$ and $r_{dB}$ equal to $r_{bB}\,[(1 - r_{ab})/2]^{1/2}$, giving the same value for $(MPP)_t$ as is found in the paragraph above where Brogden's assumptions are fully met. It is easily verified that this equation provides the same assumptions and results as does Brogden's model (1959) for his *no* selection, two-job case.

Thus it is seen that when the assignment composites have equal variances, as is true of the operational assignment/classification systems for all the military services, the contribution of HC effects vanishes and the same CE exists in our two job examples for two validities of 0.6 and 0.4 as for two equal validities of 0.5 and 0.5; they have the same value for $(MPP)_t$.

Using the relationship $r_{aA} = S_a$, $r_{bB} = S_b$, $r_{aB} = r_{ab}\, r_{bB}$, and $r_{bA} = r_{aB}\ \ r_{aA}$--all a result of defining both $a$ and $b$ as LSEs of $A$ and $B$ respectively--our formulae for $r_{dA}$ and $r_{dB}$ simplify to the following:

$$r_{dA} = (S_a{}^2 - r_{ab}\, S_a\, S_b)/(S_a{}^2 + S_b{}^2 - 2\, r_{ab}\, S_a\, S_b)^{1/2} \quad . \tag{6}$$

$$-r_{dB} = (S_b{}^2 - r_{ab}\, S_a\, S_b)/(S_a{}^2 + S_b{}^2 - 2\, r_{ab}\, S_a\, S_b)^{1/2} \quad . \tag{7}$$

When $r_{ab}$ becomes less than $(S_b/S_a)$ the sign of $r_{dB}$ changes to negative. As $r_{ab}$ approaches 1.0, the value of $r_{dA}$ approaches $S_a$ and $(-r_{dB})$ approaches $r_{bB}$. It is easily seen that this model simplifies to the model described in Appendix 1A when $r_{ab} = 1.0$.

Twelve examples in which: (1) SR = 1, (2) validities are equal to 0.5 and 0.5 or 0.6 and 0.4 for $A$ and $B$, respectively, and (3) $r_{ab}$ ranges from 0 to 1.0, are described and results in terms of MPP provided in Table 1B.2.1

**Table 1B.2.1. Demonstration of the Four Variable Model (Twelve Examples with SR = 1.0)**

| Example | $r_{ab}$ | $S_a$ | $S_b$ | $r_{dA}$ | $-r_{dB}$ | MPP |
|---|---|---|---|---|---|---|
| 1 | 0.0 | 0.5 | 0.5 | 0.35355 | 0.35355 | 0.282 |
| 2 | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 | 0.199 |
| 3 | 0.7 | 0.5 | 0.5 | 0.193649 | 0.193649 | 0.154 |
| 4 | 0.8 | 0.5 | 0.5 | 0.158114 | 0.158114 | 0.126 |
| 5 | 0.9 | 0.5 | 0.5 | 0.111803 | 0.111803 | 0.089 |
| 6 | 1.0 | 0.5 | 0.5 | 0.0 | 0.0 | 0.0 |
| 7 | 0.0 | 0.6 | 0.4 | 0.4992 | − 0.2219 | 0.288 |
| 8 | 0.5 | 0.6 | 0.4 | 0.4536 | − 0.0756 | 0.211 |
| 9 | 0.7 | 0.6 | 0.4 | 0.4476 | + 0.0187 | 0.171 |
| 10 | 0.8 | 0.6 | 0.4 | 0.4556 | + 0.0868 | 0.147 |
| 11 | 0.9 | 0.6 | 0.4 | 0.4854 | + 0.1888 | 0.118 |
| 12 | 1.0 | 0.6 | 0.4 | 0.60 | + 0.40 | 0.080 |

Although the MPP values in Table 1B.2.1 were computed using our four variable model, an entirely different approach than was used by Brogden (1959), the results for examples 1 through 6 that meet Brogden's assumptions of equal validities are precisely the same as his. Similarly, examples 6 and 12, which can be solved using the model of Appendix 1A, yield the same results across the two models. Of our twelve examples, only 7 through 11 require the more complex model of Appendix 1B.

We pay particular attention to example 10, since a key example with an SR of 0.7 described in the next section has validities of 0.6 and 0.4 with $r_{ab}$ equal to 0.8. This example is used to illustrate the use of the four variable model in a two-stage selection-classification situation. Note that here, with no selection and $r_{ab} = 0.8$, an MPP of 0.147 is achieved with hierarchical classification (HC) effects present. This value is reduced to 0.126 when the HC effects are not allowed to function, as when operational test composites are given equal SDs. This is a reduction of 14.3 percent due to elimination of HC effects. When $r_{ab} = 0$, this reduction is much less--the difference between 0.288 and 0.282--a 2.1 percent reduction.

The allocation and HC processes are clearly not additive, but are instead competitive. In this competition HC becomes predominant as $r_{ab}$ approaches 1.0 and allocation becomes predominant as $r_{ab}$ approaches zero. The effect of allocation in competition with HC intuitively appears to strengthen as the number of jobs is increased. As noted before, adding jobs over a minimum of two contributes little or nothing to the HC effects on MPP. On the other hand, adding jobs has a major positive effect on MPP in the allocation situation. We see below that allocation is also strengthened when selection is introduced.

## APPENDIX 1B.3: DEMONSTRATING THE MODEL WITH SELECTION; THE FOUR VARIABLE MODEL WITH SR = 0.7

The application of the four variable model to hypothetical operational situations in which SR and $r_{ab}$ are both less than 1.0 requires the use of variables corrected for the selection effects of a single predictor variable (referred to here as $g$). We assume a normally distributed $g$, along with $a$ and $b$, with means of 0 and SDs of 1.0 in the population from which selection is accomplished. Correlation coefficients and SDs corrected to reflect the effects of selection on $g$ are written in bold face and underlined.

Our four variable model written in its most general form includes the element $M_p$, which as the mean of the selected group was zero, and thus ignored, in our previous 12 examples. Also, both $S_A$ and $S_B$ are equal to 1.0 when SR = 1.0 and thus did not need to be written explicitly as a multiplier in our computing formulae. Our basic model for use when SR < 1.0 is as follows:

$$(MPP)_A = \underline{r}_{dA}\underline{S}_A\, M_{qA} + r_{gA}\, M_p \quad . \tag{8}$$

$$(MPP)_B = (-\underline{r}_{dB})\, \underline{S}_B\, M_{qB} + r_{gB}\, M_p \quad . \tag{9}$$

Since we define our examples as having equal quotas for each job, $M_{qA} = 0.7978$ and $M_{qB} = -0.7978$. The value of $M_p$, as the mean of the upper 70 percent of a population for a normal deviate with a mean of zero and an SD of 1.0, is 0.49673. We also require the SD of a truncated normal deviate reflecting the selection effects of an SR of 0.7. The standard deviation of this truncated normal deviate will be designated as $\underline{S}_g$.

We obtain $\underline{S}_q$ by integrating the normal density function, using integration by parts, giving us the following computing formula:

$$\underline{S}_q^{\,2} = (xz/p) - (z/p)^2 + 1 \quad , \tag{10}$$

where p equals the SR (.7 for our examples), x is the abscissa at the point that cuts off the lower 30 percent of the applicant population, and z is the ordinate of the normal curve at this same point. For our examples in which p = 0.7, x = 0.52441, and z = 0.34771, $\underline{S}_g = 0.49277$.

Our first example is for a situation in which a single variable, $g$, is used to select and assign personnel to both of two jobs, $A$ and $B$; $r_{gA} = 0.6$, $r_{gB} = 0.4$. We adjust $r_{gA}$ and $r_{gB}$ to effect a restriction in range on $g$ using the two formule given below:

$$(\underline{r}_{gA})^2 = (r_{gA}^{\,2}\, \underline{S}_g^{\,2})/[1 + r_{gA}^{\,2}\, (\underline{S}_g^{\,2} - 1)] \tag{11}$$

$$(\underline{r}_{gA})^2 = 0.217028$$

$$(\underline{r}_{gA})^2 = (r_{gB}^{\,2}\, \underline{S}_g^{\,2})/[1 + r_{gB}^{\,2}\, (\underline{S}_g^{\,2} - 1)] \tag{12}$$

$$(\underline{r}_{gA})^2 = 0.085808$$

We obtain $\underline{S}_A$ and $\underline{S}_B$ using the following formlae:

$$\underline{S}_A^{\,2} = (1 - r_{gA}^{\,2})/(1 - \underline{r}_{gA}^{\,2}) \quad (13)$$

$$\underline{S}_A = 0.9041$$

$$\underline{S}_B^{\,2} = (1 - r_{gB}^{\,2})/(1 - \underline{r}_{gA}^{\,2}) \quad (14)$$

$$\underline{S}_B = 0.95856 \quad .$$

Our values for $\underline{r}_{dA}$ and $\underline{r}_{dB}$ resulting from the use of our general formulae provided above are equal to $\underline{r}_{gA}$ and $\underline{r}_{gB}$, respectively, when SR = 1.0. Thus $(MPP)_A$ and $(MPP)_B$ can be computed from the information provided above, using formulae 8 and 9, as follows:

$$(MPP)_A = (0.46586)(0.9041)(+\,0.7978) + (0.60)(0.49673)$$

$$(MPP)_B = (0.29293)(0.9586)(-\,0.7978) + (0.40)(0.49673) \quad .$$

Thus $(MPP)_A$ equals 0.6341 and $(MPP)_B$ equals – 0.0253,and $(MPP)_t$ equals *the average of* $(MPP)_A$ and $(MPP)_B$, or 0.3044. The selection effect is equal to 0.248 and the hierarchical classification effect is equal to 0.056. The allocation effect is, of course, zero.

The above example illustrates the four-variable model with an example possessing an SR less than 1.0 and that also has $r_{ab}$ equal to 1.0--permitting confirmation by the more simple model described in Appendix 1A. The results are the same. We will now proceed to two examples that require the complexity of this more general model.

Our second example with an SR of 0.7 has allocation effects but no HC effects. For this example, $r_{ab} = 0.8$, and $r_{aA} = r_{bB} = 0.5$. Selection is accomplished on $g = a + b$. Since $S_a = S_b = 0.5$, we see that $S_g{}^2 = 0.9$, and $r_{ag} = r_{bg} = S_g = 0.4743$.

Our model requires that we have ($\underline{r}_{aA} = \underline{r}_{bB}$) and $\underline{r}_{aB}$ in order to compute $\underline{r}_{dA}$, $\underline{r}_{da}$ = (– $\underline{r}_{dB}$). We also require $\underline{r}_{gA}{}^2$,$\underline{r}_{gA}{}^2 = \underline{r}_{gB}{}^2$, to enable the computation of $\underline{S}_A$. Our reverse restriction in range formulae used to obtain these values are given below:

$$\underline{r}_{aA} = (r_{aA} + r_{gA}\, r_{gA}\, (\underline{S}_g - 1))/** \tag{15}$$

$$** = ((\,1 + r_{ga}{}^2\, (\underline{S}_g - 1))(1 + r_{gA}{}^2\, (\underline{S}_g - 1)))^2$$

$$\underline{r}_{aA} = 0.3916$$

$$\underline{r}_{ab} = (r_{ab} + r_{ga}\, r_{gb}\, (\underline{S}_g - 1))/*** \tag{16}$$

$$*** = ((\,1 + r_{ga}{}^2\, (\underline{S}_g - 1))(1 + r_{gb}{}^2\, (\underline{S}_g - 1)))^2$$

$$\underline{r}_{ab} = 0.6320$$

$$\underline{r}_{gA}{}^2 = (r_{gA}{}^2\, \underline{S}_g{}^2)/(1 + r_{gA}{}^2\, (\underline{S}_g{}^2 - 1)) \tag{17}$$

$$\underline{r}_{gA}{}^2 = 0.12516$$

$$\underline{S}_A{}^2 = (1 - {}_{gA}{}^2)/(1 - \underline{r}_{gA}{}^2) \tag{18}$$

$$\underline{S}_g = 0.9412 \quad .$$

Our computing formulae for $\underline{r}_{dA}$ can be simplified because $r_{aA} = r_{bB}$, in accordance with an explanation in a previous section. Our formula and result becomes:

$$\underline{r}_{dA} = \underline{r}_{aA}\,((1 - \underline{r}_{ab})/2)^{1/2} = 0.167993 \quad .$$

We insert these values in our basic model (formula 8) as below:

$$\begin{aligned}
(MPP)_A &= \underline{r}_{dA}\,\underline{S}_A\,M_{qA} + r_{gA}\,M_q \\
&= (0.1680)(0.9412)(0.7978) + (0.47434)(0.49673) \\
&= 0.1261 + 0.2356 \\
&= 0.3617 \quad .
\end{aligned}$$

For this example, $(MPP)_A = (MPP)_B = (MPP)_t$, and all of the classification effects are due to pure allocation. The gain in MPP over chance selection and classification due to allocation effects is 0.1261, and the comparable gain due to selection effects is 0.2356. While $(MPP)_t$ is less than is present in our next example that has disparate validities, 0.4 and 0.6 with an average validity of 0.5, the MPP of one job is not magnified at the expense of the others. Quality is level across the two jobs, a goal frequently pursued by military managers.

Our third example with an SR of 0.7 has both allocation and HC effects. The $r_{ab}$ remains at 0.8, but $r_{aA} = 0.6$ and $r_{bB} = 0.4$. Selection is still accomplished on $g = a + b$, but $g$ is a different variable since $a$ and $b$ no longer have equal SDs--instead their SDs are respectively 0.6 and 0.8.

Using "correlation of sums" formulae we see that $S_{g2} = 0.904$, $r_{ag} = 0.9676$, $r_{bg} = 0.9255$, $r_{gA} = 0.5806$, $r_{gB} = 0.3702$. The same general formulae as in 15 through 18 above, are used to make reverse restriction in range corrections providing the following results for the indicated relationships.

$$\begin{aligned}
&\underline{r}_{ab} = 0.6345 \\
&\underline{r}_{aA} = 0.4775\ ;\ \underline{r}_{bA} = 0.3030 \\
&\underline{r}_{bB} = 0.31183;\ \underline{r}_{aB} = 0.19785 \\
&\underline{r}_{ag}{}^2 = 0.87864;\ \underline{r}_{bg}{}^2 = 0.74635 \\
&\underline{S}_A = 0.4348\ ;\ \underline{S}_b = 0.3008 \\
&\underline{S}_A = 0.9105\ ;\ \underline{S}_B = 0.9646 \quad .
\end{aligned}$$

We must use the more general formula for computing $\underline{r}_{dA}$ and $(-\underline{r}_{dB})$, since $\underline{S}_a$ does not equal $\underline{S}_b$ in the restricted space, although $\underline{r}_{bA} = \underline{r}_{aA}\, \underline{r}_{ab}$--just as it does in the population. The formulae used are given below:

$$\underline{r}_{dA} = (\underline{r}_{aA}\,(\underline{S}_a - \underline{r}_{ab}\,\underline{S}_b))/\underline{S}_d \tag{19}$$

$$\underline{r}_{dA} = 0.116478/.33698 = 0.3456$$

$$-\underline{r}_{dB} = (\underline{r}_{aB}\,(\underline{S}_b - \underline{r}_{ab}\,\underline{S}_a))/\underline{S}_d \tag{20}$$

$$-\underline{r}_{dB} = -\,0.007776/0.33698 = -\,0.0231$$

$$\underline{S}_d = \underline{S}_a^2 + \underline{S}_a^2 - 2\,\underline{r}_{ab}\,\underline{S}_a\,\underline{S}_b = 0\ 33698 \quad .$$

Inserting the appropriate values from above into our basic model (formulae 8 and 9) provides the following results:

$$\begin{aligned}
(MPP)_A &= \underline{r}_{dA}\,\underline{S}_A\, M_{qA} + r_{gA}\, M_p \\
&= (0.3456)(0.9105)(0.7978) + (0.5806)(0.49673) \\
&= 0.2510 + 0.2884 \\
&= 0.5394
\end{aligned}$$

$$\begin{aligned}
(MPP)_B &= \underline{r}_{dB}\,\underline{S}_B\, M_{qB} + r_{gA}\, M_p;\ M_{qB} = (-M_{qA}) \text{ when } q = 0.5 \\
&= (-\,0.0231)(-\,0.7978)(0.9646) + (0.3702)(0.49673) \\
&= 0.0178 + 0.1839 \\
&= 0.2017
\end{aligned}$$

$$(MPP)_t = ((MPP)_A + (MPP)_B)/2 = 0.3706 \quad .$$

In this last example the gain in MPP over chance selection and assignment due to classification effects (both HC and allocation) is 0.1344 and the comparable gain due to selection effects is 0.2362. Comparing the results of the two examples, we see that the loss of classification efficiency (measured in terms of MPP) due to elimination of HC effects (e.g., by transforming a and b scores so as to give them equal variances in an operational situation) results in a loss of 0.008 of MPP measured in standard scores, a 6 percent loss. This compares with a loss of 14.3 percent for a comparable situation without selection. The results for the three examples with SR = 0.7 are summarized in Table 1B.3.1.

**Table 1B.3.1. Three Examples With SR = 0.7**
**(Entries are MPP Standard Scores)**

| Example # | Total Gain | Gain Due to Selection | Source of Gain Due to Classification | Classification Efficiency |
|---|---|---|---|---|
| 1 | 0.304 | 0.248 | 0.056 | HC Only |
| 2 | 0.362 | 0.236 | 0.126 | Allocation Only |
| 3 | 0.371 | 0.237 | 0.134 | HC and Allocation |

Our four-variable model can be used to evaluate a variety of operational conditions involving selection and/or optimal assignment to two jobs. Actual empirical data for two predictors and two criteria variables can be used in conjunction with the model in its most general form. Predictor composites need not be LSEs; service aptitude area composites can be depicted in our model.

We chose to illustrate our model with examples that either permitted comparisons with Brogden's model (1959), or provided a means of comparing pure HC, pure allocation, or mixed situations--with hypothetical relationships among the variables that fall within a range that is frequently encountered in real life. However, this model can be readily utilized to investigate other issues, such as the effect of validity range and magnitude of average validities on HC and mixed HC and allocation situations.

Consideration of the results for our selected examples has sharpened our intuition with respect to the competitive relationship between HC and allocation effects. The competitive position of allocation with respect to HC is greatly increased as more jobs and and corresponding test composites are added to the classification system. The reverse is true with respect to the average intercorrelation coefficients among LSEs used as predictors. As r approaches 1.0 the role of allocation literally vanishes, while the competitive role of HC becomes trivial when *r* becomes small (a small r is a rather unlikely finding in real life). All things considered, the elimination of HC effects from the operational test composites used for personnel classification in the services may not have as much adverse impact on the magnitude of MPP as we initially thought.

Simulation provides a more precise method for investigating such issues when there are several composites and job families. One kind of simulation methodology appropriate for this purpose is discussed in Chapter 4, and another kind is illustrated by Nord and Schmitz in Chapter 3 of Zeidner and Johnson (1989).

# CHAPTER 2. MEASUREMENT OF CLASSIFICATION EFFECTIVENESS

## A. INTRODUCTION

The work of Brogden (1946b, 1951, 1954a, 1954b, 1955, 1959, 1964) and Horst (1954, 1956a, 1956b, 1960a, 1960b) generated the main stream of progress in the measurement and improvement of classification effectiveness. Brogden directly ties measurement of classification efficiency to mean predicted performance (MPP) and thus to utility. Horst's measure of classification efficiency has a direct and simple relationship to Brogden's measure; the square root of Horst's index may be adjusted to make it proportional to Brogden's when the same assumptions are made. This adjusted index thus measures the benefit obtainable from a classification test battery for a specified set of jobs (i.e., PCE). This is especially fortunate since Horst's index has a number of advantages: it is simple to compute and to adjust; it is readily adapted for use in selecting tests for inclusion in a classification battery; and it may provide more robust estimates than Brogden's measures with departures from assumptions.

Mean predicted performance (MPP), used by Brogden as the measure of both operational effectiveness and potential efficiency of selection/classification, is the same measure Brogden used in unidimensional selection. Brogden's historic contribution wherein he used correlation coefficients as least square regression weights to provide MPP measures, led naturally to the expression of classification in the same terms. Additionally, an improvement in the selection ratio was seen by Brogden to produce similar benefits for a least squares weighted prediction estimate (LSE) computed separately for each job as produced in unidimensional selection (Brogden, 1959).

It is difficult to envisage the use of several *different* LSEs (each corresponding to a different job) to select from a common applicant pool without also stipulating an assignment algorithm. Multidimensional selection is maximally effective when the LSE score corresponding to the job to which an individual has been assigned is higher than any score for the same LSE in the rejected group. In order to make reject/accept decisions, the scores of all applicants across all jobs must be compared. The applicant cannot simply be rank ordered only on his highest LSE score; quotas may impact on the assignment process in such a way as to force consideration of an applicant's acceptance in a different job based on his rank order with respect to that job's corresponding LSE score.

It is not difficult to envisage a multidimensional assignment process which simultaneously maximizes selection and classification. Such a process was indicated by Brogden (1951, 1959). Brogden developed as the source of the gains attributed to the joint application of selection and classification procedures a model in which every applicant, if selected, was assigned to the job corresponding to that LSE score with the highest unique component; the applicant was not selected if, and only if, he had no unique component score as high as the highest unique component score of an accepted applicant. Selection is on the unique components ("$u$") that are necessary and sufficient for classification. In this selection on "$u$" model, the MPP in the non-selected group was reduced, but not minimized since selection was not accomplished on the total LSE score. However, in Brogden's multidimensional selection model, the increase in the MPP standard score resulting from assignment, as compared to random assignment, was maximized in the accepted group. While assignment does not suffer a loss in efficiency from the exclusive use of "$u$" in the selection-assignment process, selection clearly does.

An operational selection and assignment algorithm for implementing this model is impractical. Since the effect of the general component is not considered in the selection process, the model does not accomplish selection with a set of variables that would ordinarily be used operationally. However, this model reflects the potential utilization efficiency obtainable under Brogden's assumptions, including his selection classification process,[5] for a defined test battery and set of jobs; and very importantly, Brogden's measure of classification efficiency is readily convertible to utility terms.

This chapter focuses on the measurement of potential classification efficiency in terms of MPP; other measures of classification effectiveness are discussed in Zeidner and Johnson (1989b). The related contributions of Brogden and Horst to the increase of MPP by selecting efficient classification tests for inclusion in a test battery, by selecting more efficient test composites, and/or by restructuring job families, will be discussed in the following chapter. Brogden's and Horst's contributions to the improvement of

---

5 Brogden's selection-classification process is unfortunately not usually listed as an assumption; his results listed in Table 1 (1959) depend upon his particular selection-classification process and this process is a key assumption of his model.

classification efficiency are described separately to highlight methodological issues related to the increasing of potential classification efficiency (PCE).

## B. BROGDEN'S CONTRIBUTION TO THE MEASUREMENT OF POTENTIAL UTILIZATION EFFICIENCY

To explore the benefits of differential selection and allocation of applicants, Brogden (1951) applied the approach he had previously used (1946a) to measure mean predicted performance (MPP) in the unidimensional selection case. To this end, he introduced the concept of a differential selection model whose implementation has been referred to in the previous chapter as the "multidimensional screening" (MDS) algorithm. Benefits that could be provided by multidimensional selection and classification were not limited to, but were primarily thought of, initially, in terms of an improvement of the selection ratio. For example, in selecting for two jobs using separate predictors correlating less than 1.0 with each other, Brogden (1951) noted that since each predictor was a "composite derived by multiple correlation procedures...."[6] (p. 176), a higher cut score on each predictor would yield the same number of qualified, selected applicants as a lower cut score on a univariate selector, the higher cut scores, of course, yielding higher MPP standard scores. Since some applicants would be rejected by both predictors, the improvement in the selection ratio would be a function of this overlap. The full potential of this increase in the MPP of selected applicants would be realizable only if an optimal assignment process were used to allocate successful applicants. However, if, after the rejection/acceptance decision were made using one LSE per job, employees were assigned randomly (but only among those jobs for which they exceed the cutting score on the corresponding LSE) the advantage of the improved selection ratio would be partially maintained while the gain from optimal assignment would be minimized. The effects on MPP of improving the selection ratio and from optimal asignment could be partially separated in this fashion,

Table 1 of Brogden's 1951 article showed MPP in terms of standard scores for selection situations involving differential assignment to two jobs and percents rejected ranging from 10 percent to 90 percent, the correlation between the two predictors ranging from zero to 1.0, and validities equal to 0.5. Brogden provides a footnote explaining the MPP values corresponding to an intercorrelation of 1.0. (See Table 2.1.) He notes,

---

6 It should be noted that the standard deviation of each predictor is defined in the appendix of Brogden's article as the multiple correlation coefficient, thus identifying the "predictor" as necessarily a LSE.

"Assignment with two predictors correlating unity is equivalent to assignment with a single predictor." (p. 182, Table 2). This is, of course, correct but seems inconsistent with his 1959 article, particularly tne values in his Table 1 and the procedure he recommends of multiplying table entries by $R\sqrt{1-r}$. We believe the apparent inconsistency is attributable to a different selection process underlying the models used in the two articles, selection on the LSEs as contrasted to selection on the unique components.

**Table 2.1. Mean Standard Criterion Values Resulting from Differential Placement Into Two Assignments as a Function of the Degree of Correlation Between the two Predictors* and the Percentage Placed In Each Assignment**

| Percentage Placed in Each of Two Assignments | Correlation Between Predictors | | | | | |
|---|---|---|---|---|---|---|
| | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0** |
| 5% | 1.03 | 1.02 | 1.01 | 1.00 | 0.96 | 0.88 |
| 10% | 0.87 | 0.86 | 0.84 | 0.82 | 0.79 | 0.70 |
| 15% | 0.76 | 0.75 | 0.73 | 0.71 | 0.68 | 0.58 |
| 20% | 0.68 | 0.67 | 0.65 | 0.62 | 0.59 | 0.48 |
| 25% | 0.61 | 0.60 | 0.57 | 0.54 | 0.51 | 0.40 |
| 30% | 0.55 | 0.53 | 0.50 | 0.46 | 0.43 | 0.32 |
| 35% | 0.48 | 0.47 | 0.43 | 0.40 | 0.36 | 0.25 |
| 40% | 0.42 | 0.41 | 0.37 | 0.34 | 0.29 | 0.18 |
| 45% | 0.36 | 0.34 | 0.30 | 0.26 | 0.22 | 0.10 |
| 50% | 0.31 | 0.28 | 0.25 | 0.22 | 0.17 | 0.00 |

Source: Brogden (1951), p. 182.

* Each of the two predictors is assumed to have a validity of 0.5.

** Assignment with two predictors correlating unity is equivalent to assignment with a single predictor.

The more general solution for the value of MPP provided by Brogden (1959) was based on: (1) LSE intercorrelations; (2) number of jobs; (3) the value of a common validity for all jobs; and (4) the percent rejected. Again, a model is provided for the selection and allocation of applicants using an algorithm equivalent to the MDS, but visualized as applying only to the unique component of the LSEs. For this solution, Brogden uses

Tippet's (1925) tables to arrive at MPP values for zero correlated predictors assumed to have validity coefficients equal to 1.0. Actually, Brogden's values shown in the table are criterion standard scores which can be converted to MPP standard scores, as previously noted, by multiplying the mean criterion scores by the validities of the LSEs. The mean criterion scores derived from Brogden's table will be referred to here as $M_{p,m}$, where the subscript $p$ denotes the percent of the applicant population rejected, and the subscript $m$ denotes the number of jobs among which selected applicants are optimally allocated.

The variables referred to as predictors by Brogden are defined precisely by him as least squares weighted performance estimates (LSEs), where the separate regression equation for each job is based on all variables in a battery of predictor measures. As previously noted in Chapter 1, those LSEs are optimal for classification as well as for selection.

A provision for correlated predictors in a model based on orthogonal and unique components is made by means of an assumption that the magnitude of the intercorrelation among LSEs is attributable to the presence of an underlying general component, "g." All remaining reliable variance is attributable to a set of unique components. These unique components corresponding to each job are uncorrelated with each other and with "g;" each such factor, referred to as "u," has a common (i.e., equal among all "u" variables) validity value with its corresponding criterion and a zero relationship with all others. Brogden made an analogy between these particular assumptions and the concept of parallel form tests in which each test consists of a true score and an error component that is uncorrelated with both the true score and the error components in other predictors. The standard deviations of the remaining orthogonal components of the several predictors, after $g$ has been removed, are shown by means of the above analogy to be $\sqrt{1-r}$ , where $r$ is the value of all of the intercorrelations among the predictors.[7]

The value of $\sqrt{1-r}$ is used as a multiplier to scale the tabled values of $M_{p,m}$ to provide an MPP standard score for correlated predictors (LSEs). The value of $\sqrt{1-r}\ M_{p,m}$ is the MPP standard score for perfectly valid predictors (i.e., the criterion variables); more generally, $R\sqrt{1-r}\ M_{p,m}$ is the MPP standard score resulting from Brogden's selection and classification process. This is a measure of PAE when there is no selection; otherwise it is

[7] We use an alternative to this analogy later on in this chapter; our four-variable model described in Appendix 1B confirms Brogden's results for the two-predictor case.

an underestimate of the PUE that would result from using the LSEs in the selection as well as the assignment process.

If each LSE score is separated into a unique score called $u$ and a common or general score, $g$, corresponding to the $g$ and $u$ components described above, we can say that the predicted performance score, $y$, equals the weighted sum of $u$ and $g$ expressed as standard scores, as if they were factor scores in a factor model. Brogden points out that only the $u$ component contributes to allocation efficiency, and he makes no use of the $g$ component in computing his table entries ($M_{p,m}$) except when $m$ equals one, the simple univariate selection case. There is, of course, no distinction between $g$ and $u$ when $m$ equals one. Brogden's use of a multiplier equal to $R\sqrt{1\text{-}r}$ when $r = 1.0$ reduces the value of the MPP standard score to zero in situations where PUE (the MPP standard score obtained after using an optimal selection-classification process) cannot be less than $R\ (M_{1,m,})$, if selection is accomplished on the LSEs instead of on the $u$ values.

The Brogden model is not representative of most operational situations when $p > 0$, since $g$ is not used in the selection process, although the value of $g$ (e.g., general mental ability) for this process is universally recognized. In fact, the selection effects in almost any operational process should yield an MPP standard score as large as that provided by a single predictor when $r = 1.0$, regardless of the value of $m$. Thus the entries in Brogden's Table 1 (1959, p. 189) provide correct PUE values only for the row corresponding to $m = 1$ (i.e., for selection to one job across all values of $p$), and for the column corresponding to $p = 0$ (i.e., for no rejectees). (See Table 2.2.) We will refer to all other values derived from his table and multiplied by $R\sqrt{1\text{-}r}$ as estimates, rather than measures, of PUE.

The selection process which would correspond to Brogden's tabled MMP standard scores has an effect equivalent to our MDS algorithm, except that Brogden's model uses only the unique components of the LSEs, rather than the total LSE scores as the selection and assignment variables. While the same classification results would be obtained using either the total LSE scores or the unique components of these scores, the same is not true for the selection process. The general factor components of the LSEs can make a major contribution to PSE, in addition to the contribution that unique scores can make. Thus Brogden's tabled MPP values are correct for the implied selection process, but it is fairly unlikely that this particular process, one in which the estimated PUE is zero for a unidimensional battery used for both selection and assignment, will ever be used in an operational situation.

**Table 2.2. The Allocation Average as a Function of Percent of the Applicant Pool Rejected and the Number of Jobs**

| Number of Jobs | When $R$ = 1.00 and $r$ = 0* % of $N$ Rejected | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| 1 | 0.00 | 0.20 | 0.35 | 0.50 | 0.64 | 0.80 | 0.97 | 1.16 | 1.40 | 1.75 |
| 2 | 0.56 | 0.73 | 0.85 | 0.97 | 1.09 | 1.22 | 1.37 | 1.54 | 1.75 | 2.07 |
| 3 | 0.85 | 0.99 | 1.10 | 1.21 | 1.32 | 1.44 | 1.57 | 1.73 | 1.93 | 2.23 |
| 4 | 1.03 | 1.17 | 1.27 | 1.37 | 1.48 | 1.59 | 1.71 | 1.86 | 2.05 | 2.35 |
| 5 | 1.16 | 1.29 | 1.39 | 1.49 | 1.59 | 1.70 | 1.82 | 1.95 | 2.14 | 2.43 |
| 6 | 1.27 | 1.38 | 1.48 | 1.58 | 1.68 | 1.78 | 1.90 | 2.04 | 2.22 | 2.51 |
| 7 | 1.35 | 1.46 | 1.56 | 1.65 | 1.75 | 1.86 | 1.97 | 2.10 | 2.28 | 2.55 |
| 8 | 1.42 | 1.53 | 1.63 | 1.72 | 1.81 | 1.91 | 2.03 | 2.16 | 2.33 | 2.60 |
| 9 | 1.49 | 1.59 | 1.68 | 1.77 | 1.86 | 1.96 | 2.07 | 2.20 | 2.38 | 2.64 |
| 10 | 1.54 | 1.65 | 1.73 | 1.82 | 1.91 | 2.01 | 2.11 | 2.24 | 2.41 | 2.68 |

Source: Brogden (1959), p. 189.

* To calculate an allocation average for other specified values of $R$ (the validity of the $\hat{C}_{ij}$) and $r$ (the intercorrelation of the $\hat{C}_{ij}$), multiply by $R\sqrt{1-r}$ .

We develop and describe two modifications of Brogden's (1959) model; each incorporates one of two alternative selection processes. Both modifications use Brogden's tabled values as a starting point. Both models provide the same results as Brogden's when no one is rejected, and, for the "selection on $g$ and $u$" model, the same results are provided when $r$ equals zero. For both modifications, (selection on $g$ and selection on $u$ and $g$), when $r$ is equal to unity, the MPP standard score will be the same as when $m$ equals *one* (the univariate case). The latter desirable relationship does not hold for Brogden's model (i.e., the selection on "$u$" model).

The first of these two modifications uses a selection process analogous to the two-stage selection procedure in which selection is accomplished using the $g$ component and classification using the $u$ component of the LSEs. This process resembles the most commonly used selection/classification process in which the rejection/acceptance decision is made on general mental ability and the later classification process is accomplished using more job specific measures. The rejection of applicants with the lowest $g$ component scores over the total applicant group does not affect the mean of the unique component for those accepted and assigned to jobs, since $u$ and $g$ are independent of each other.

The second model uses cut scores on both $g$ and $u_j$ to effect selection. The rejection of 10% ($p = 0.10$) on each of the $g$ and $u$ components of LSE for each job will yield a selection ratio of $(1\text{-}p)^2$ or 0.81; a separate $p$ on both $g$ and each of the $m$ $u_j$ of 0.20 is equivalent to a SR of 0.64, and a separate $p$ on both $g$ and each of the $m$ $u_j$ of 0.30 is equivalent to a SR of 0.49.

A formula for using Brogden's table of $M_{p,m}$ values to compute MPP standard score values corresponding to each of these modifications is derived using Brogden's assumptions, which require the covariance matrix among the predictors (i.e., predicted performance estimates, LSEs) to have the same value for all diagonal elements, $R^2$, and a different common value for all of the off diagonal elements ($R^2r$ ). The corresponding intercorrelation matrix has diagonal elements of unity and off diagonal elements of $r$. Also each LSE score consists of a $g$ component which is the same for all LSE scores belonging to a given individual, and a separate $u$ component for each LSE score that is uncorrelated with the $g$ component, or with the $u$ components of the LSEs of other jobs.

It is useful for the development of formulas for MPP based on modifications of Brogden's model (1959) and for comparison of Brogden's and Horst's measures of classification efficiency, to express Brogden's (1959) assumptions in the form of a particular factor extension matrix, **F**. This particular **F** matrix reproduces the covariance matrix among the LSEs (see Appendix 2A). Expressed as a general matrix formula, **FF'** = **C**, each row of **F** represents a LSE for a particular job; the columns represent factors--one general factor and $m$ unique factors that have only one non-zero element in each column. In the special set of values for **F** that represents Brogden's assumptions, each element (factor coefficient) of the general factor, $g$, has the value of $R\sqrt{r}$ , while each non-zero factor coefficient of the $m$ unique factors has a coefficient of $R\sqrt{1-r}$ . If these values for the elements of the particular **F** that expresses Brogden's assumptions are used, the assumed values for the elements of **C** that fulfill Brogden's assumptions are readily reproduced, **FF'** = **C**.

Since **F** is an orthogonal factor solution, $g$ and each of the unique factors, $u_1$, $u_2$ ... $u_m$, represent variables having a mean of zero and a standard deviation of one; these $m + 1$ column variables are mutually uncorrelated. The elements of **F** are factor coefficients, sometimes called factor loadings, which are both (1) the correlations between the LSEs for each job and the factors, $g$ and $u_j$ and (2) the regression weights that can be applied to the factor scores to form $m$ regression equations in which the LSE scores are the dependent variables. Thus for the $i^{th}$ individual assigned to the $j^{th}$ job, the LSE score referred to as $y_{ij}$, can be expressed as $y_{ij} = R\sqrt{r}\ \ g_i + R\sqrt{1-r}\ \ u_{ij}$ .

It is well known that the correlations of $g$ and $u_j$ with the $j^{th}$ LSE are equal to the correlations of $g$ and $u$ with the actual criterion scores for the $j^{th}$ job. Thus the regression weights for predicting LSE are the same regression weights as predict the performance criterion scores. Mean predicted performance on the $j^{th}$ job, for any defined subgroup, is the sum of the mean $g$ score and the mean $u_j$ score for the individuals assigned to that group. Since our assumptions are the same as Brogden's, the same number of individuals, $N_p$, are assigned to each job, $m(N_p)$ individuals are assigned to all $m$ jobs. The mean predicted performance standard score for the $j^{th}$ job, when a specified percentage, $p$, is rejected using the two stage selection/classification procedure (i.e., the use of $g$, rather than $u$ in the selection procedure) will be denoted as $(\mathrm{MPP})_{pj}$. Thus the following formula holds[8] :

$$(\mathrm{MPP})_{pj} = (1/N_p\ R\sqrt{r}\ \Sigma^{(1)} g_i + (1/N_p)\ R\sqrt{1-r} \left( \Sigma^{(2)} \Sigma^{(3)} u_{ij} + \Sigma^{(1)} u_{ij} \right) .$$

The summing over $i$ is accomplished on the $N_p$ largest of the $g$ scores (with $u_j$ and $g$ summed separately). The selection of the $N_p$ largest g scores is obtained by placing these scores in rank order and accepting the highest $N_p$ scores. Since the $u_j$ and $g$ variables are in standard score form, the MPP score will also be expressed in terms of a variable that has a mean of zero and a standard deviation of one--the statistical characteristics we wish MPP to have in the youth population. The following notation will be utilized:

$$M_{p,1} = 1/N_p\ \Sigma^{(1)} g_i \ \ ; \ M_{p,m} = 1/N_p\ \Sigma^{(2)} \Sigma^{(3)} u_i + \Sigma^{(1)} u_i \ \ .$$

---

[8] There are three regions over which the summing is accomplished: (1) summing over the region containing all those accepted for entry into the Army, a region that varies with the value of $p$; (2) summing over the $m$ jobs; (3) summing over those accepted and assigned to any one of $m$ jobs; under Brogden's assumptions the expected value of this sum is the same for all jobs; this region varies with the values of both $p$ and $m$. The three regions are designated by superscripts on $\Sigma$ [e.g., $\Sigma^{(1)}$].

Note that the use of $j$ as a subscript can be dropped since under Brogden's assumptions each job has exactly the same statistical characteristics; each mean criterion standard score refers to one job, but any job.

The means of the $g$ scores for the selected personnel under several different selection ratios (i.e., $p$) have been tabled by Brogden; these tabled values have been referred to above as $M_{p,1}$. Similarly the means of the $u$ scores in the applicant group corresponding to the job to which each individual would be assigned, if selected, has been designated as $M_{0,m}$, when $m > 1$, and are obtainable from Brogden's Table 1. Thus the correct MPP standard score values for 2 or more jobs, as provided by the "selection on $g$" model, can be computed as follows:

$$(\mathrm{MPP})_{pm} = R\sqrt{r}\ M_{p1} + R\sqrt{1-r}\ M_{0m}\ ; m > 1 \quad . \tag{2.1}$$

Table 2.3 containing selected values of $(1/R)(\mathrm{MPP})_{pm}$ are provided below for selected values of $r$, $p$, and $m$. The values provided in this table are criterion means and should be multiplied by $R$ to obtain values for MPP standard scores (estimated PUE). These values contrast to the multiplier of $R\sqrt{1-r}$ stipulated by Brogden for application to the values from his table. (See Table 2.2.)

We find complete agreement between Brogden's values for MPP standard scores and those computed by Equation (2.1) when $p$ equals zero. This allocation case, where $p$ equals zero, is used in the following section to establish the link between Brogden's and Horst's measures of allocation efficiency.

The values of MPP standard scores provided in Table 2.3 support the general conclusions reached by Brogden (1959) and are based on the assumptions and insights provided in his pathfinding article. MPP standard score values from this table can provide personnel management with estimates of the gains in performance realizable from the use of LSEs as aptitude area composites, given that the most efficient assignment processes were utilized after initial selection on general mental ability.

PUE values are underestimates because PSE has not been maximized for the multivariate case in either Brogden's (1959) model or in the modifications provided here. In the Army case there is a compensating effect (to some unknown degree); the PUEs are overestimates with respect to Army input because the applicant population, as contrasted with the youth population, has a skewed distribution as if censored over the upper one third

**Table 2.3. Selection-on-"*g*" Model**

| *r* | S.R. = 0.80 | | | | S.R. = 0.70 | | | |
|---|---|---|---|---|---|---|---|---|
| | *m* | | | | *m* | | | |
| | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| 1.0 | 0.35 | 0.35 | 0.35 | 0.35 | 0.50 | 0.50 | 0.50 | 0.50 |
| 0.95 | 0.47 | 0.53 | 0.57 | 0.60 | 0.61 | 0.68 | 0.72 | 0.75 |
| 0.90 | 0.50 | 0.60 | 0.66 | 0.70 | 0.65 | 0.74 | 0.80 | 0.84 |
| 0.85 | 0.54 | 0.65 | 0.72 | 0.77 | 0.69 | 0.79 | 0.86 | 0.91 |
| 0.80 | 0.56 | 0.69 | 0.77 | 0.83 | 0.70 | 0.83 | 0.91 | 0.97 |
| 0.50 | 0.64 | 0.85 | 0.98 | 1.07 | 0.75 | 0.95 | 1.08 | 1.17 |
| 0 | 0.56 | 0.85 | 1.03 | 1.16 | 0.56 | 0.85 | 1.03 | 1.16 |

| | S.R. = .60 | | | | S.R. = .50 | | | |
|---|---|---|---|---|---|---|---|---|
| | *m* | | | | *m* | | | |
| | 2 | 3 | 4 | 5 | 2 | 3 | 4 | 5 |
| 1.0 | 0.64 | 0.64 | 0.64 | 0.64 | 0.80 | 0.80 | 0.80 | 0.80 |
| 0.95 | 0.75 | 0.82 | 0.86 | 0.89 | 0.90 | 0.97 | 1.01 | 1.04 |
| 0.90 | 0.79 | 0.88 | 0.94 | 0.98 | 0.93 | 1.03 | 1.08 | 1.12 |
| 0.85 | 0.81 | 0.92 | 0.99 | 1.04 | 0.95 | 1.06 | 1.13 | 1.18 |
| 0.80 | 0.83 | 0.96 | 1.04 | 1.09 | 0.96 | 1.09 | 1.17 | 1.23 |
| 0.50 | 0.85 | 1.05 | 1.18 | 1.27 | 0.96 | 1.17 | 1.29 | 1.38 |
| 0 | 0.56 | 0.85 | 1.03 | 1.16 | 0.56 | 0.85 | 1.03 | 1.16 |

NOTE: Table values are mean criterion standard scores which become MPP standard scores when multipled by R, the common validity coefficient of the LSEs. One LSE corresponds to each job. The common intercorrelations among LSEs is represented as "*r*," and the number of jobs and the dimensionality of the joint predictor-criterion space as "*m*." All tabled values derive, after further computations, from Brogden's (1959), Table I. Assumptions are the same except for the variable on which selection is accomplished.

of the "$g$" distribution). It is, of course, important not to confuse potential efficiency with the operational efficiency obtainable from grossly imperfect selection and assignment algorithms.

The Brogden model (as well as our two modifications) will underestimate the benefits obtainable from an assignment process that capitalizes on the potential hierarchical classification efficiency present in the system. When the validities of the LSEs vary widely across jobs, the use of the mean validity to obtain an estimate of PAE from the corrected table will probably yield a reasonably accurate estimate. However, this estimated PAE will considerably underestimate the total potential classification efficiency (the combination of the potential allocation and hierarchical classification efficiency). PCE based in part on hierarchical layering effects is, of course, only realizable if the test composites are scaled so as to have means and/or variances proportional to their values and/or validities, and an optimal assignment process is utilized.

Both modifications of Brogden's models can provide values of MPP standard scores for the set of SRs tabled by Brogden; the same set of SRs, 0.10 through 0.90, could be provided for both the model for which selection is based on $g$ and the model for which selection is based on u + g. Additionally, using Brogden's tabled values, as input to our equation, can provide results for from 2 to 9 LSEs (for nine different jobs). While our two models can provide for factoring out $R$, a similar factoring out of $r$ is not feasible with respect to tabled values. Thus, for both Tables 2.3 and 2.4, the basic entries, MPP standard scores must be separately identified for each value of $r$.

For the table corresponding to the first of our two modified models, Table 2.3, we have abridged the values of $r$ provided by Brogden to seven values, 0.50, 0.80, 0.85, 0.90, 0.95, and 1.0; those ranging from 0.80 to 0.95 are within the most relevant range of values for operational batteries and situations. Similarly, we display the effects of $m$ equal to two through five because it is very unlikely that the joint predictor-criterion space will have more than five real dimensions of practical magnitude. The table entries corresponding to $r = 1$ can either relate to the situation where $m = 1$ or to a multi-job situation for which the predictor-criterion space is unidimensional. An SR of greater than 0.50 will not be used as an argument in Table 2.3 because the contribution of classification to PUE, as SR is increased beyond 0.50, becomes increasingly negligible; for higher values of SR, the contribution of selection dominates personnel utilization effects.

The modification of Brogden's model that incorporates a selection process utilizing separate and independent selection on both $u$ and $g$ creates an SR of 0.81, when the SR on each of the orthogonal components of the LSE is 0.90. Similarly, when an SR of 0.70 is applied to both components, an SR of 0.49 is provided by this model. We will refer to this model as the "selection on $u$ and $g$ model." The derivation of the formula for computing the MPP standard scores resulting from this model is provided in Appendix 2B. The equation for computing MPP standard scores resulting from selection on $g$ and $u$ and assigning on LSEs (equivalent to assigning each individual to the job corresponding to his highest $u_j$ score) is:

$$\text{Estimated } (\text{MPP})_{pm} = \sqrt{r}\ M_{p1} + \sqrt{1-r}\ M_{pm} \quad . \tag{2.2}$$

Table 2.4 permits the comparison of selected results across the three models discussed above in this section (Brogden's "selection on $u$" model, the "selection on $g$" model, and the "selection on $u$ and $g$" model). All results are based on the values provided by Brogden's 1959 Table 1 (see Table 2.2). To avoid using Tippet's (1925) data on order functions, we compare SRs of 0.81 and 0.49 for the latter of the three models with SRs of

**Table 2.4. Comparison of Three Models**

| | | Selection Process Used in Model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Using-only-"$u$" | | | Using-only-"$g$" | | | Using "$u$" and "$g$" | | |
| | | $m$ | | | $m$ | | | $m$ | | |
| $r$ | S.R. | 3 | 4 | 5 | 3 | 4 | 5 | 3 | 4 | 5 |
| 0.95 | 0.80/0.81 | 0.25 | 0.28 | 0.31 | 0.53 | 0.57 | 0.60 | 0.41 | 0.45 | 0.48 |
| 0.80 | 0.80/0.81 | 0.49 | 0.57 | 0.62 | 0.69 | 0.77 | 0.83 | 0.62 | 0.70 | 0.75 |
| 0.50 | 0.80/0.81 | 0.78 | 0.90 | 0.98 | 0.85 | 0.98 | 1.07 | 0.84 | 0.96 | 1.05 |
| 0.95 | 0.50/0.49 | 0.32 | 0.36 | 0.38 | 0.97 | 1.01 | 1.04 | 0.75 | 0.79 | 0.82 |
| 0.80 | 0.50/0.49 | 0.64 | 0.71 | 0.76 | 1.09 | 1.17 | 1.23 | 0.98 | 1.06 | 1.11 |
| 0.50 | 0.50/0.49 | 1.02 | 1.12 | 1.20 | 1.17 | 1.29 | 1.38 | 1.21 | 1.32 | 1.40 |

NOTE: All entries in the above table are mean criterion standard scores; to obtain MPP standard scores multiply these entries by $R$, the common validity of the LSEs. Entries are derived, after further computations (except for the using-only"$u$" model), from Brogden's Table 1 values (1959). All of Brogden's assumptions are also assumed in the further computations used to compute these entries.

0.80 and 0.50 respectively, for the other two models. Our table will compare the three models for each combination of these SRs with three values of $r$ (0.50, 0.80 and 0.95) and three values of $m$ (3, 4, and 5), providing 18 entries for each model.

In examining Table 2.3, one should keep in mind that selection using only one test composite, assuming random assignment to jobs after selection and an equal value for $R$ against all jobs, will provide a mean criterion standard score (the MPP standard score divided by $R$) equal to 0.35 for a SR of 0.80 and equal to 0.80 for a SR of 0.90. Any combined selection and classification process that provides lower criterion standard scores than those obtainable by selection alone for particular values of $r$, $m$, and SR is obviously ineffective. Thus, we see that a simultaneous selection-classification effort would not appear to be worthwhile using Brogden's "selection using $u$" model (as depicted in Table 2.1) for an SR of 0.80 (or more), where $r$ equals 0.90 (or more) and $m$ equals 3 (or less).

Our "selection using only $g$" modification of Brogden's model indicates higher values of PAE as a result of the two-stage selection and classification process. For SR equal to 0.80, a 51 percent gain over optimal selection, combined with random assignment to jobs, results when $r$ is equal to 0.95 and $m$ is equal to 2. For the same SR, if $r$ is lowered to 0.80 and $m$ increased to 4, a conceivable but difficult goal to achieve, the model provides a 120 percent gain over optimal selection and random assignment; for an SR of 0.50 this gain decreases to 0.46 percent.

The "selection using $u$ and $g$" model provides higher MPP scores than either of the other two selection and classification models when $r$ is equal to 0.5, but provides smaller gains, or actual decrements, as $r$ approaches either zero or unity. This model is definitely inferior to the "selection using $g$" model within the more practical range for $r$, that is, for $r$ equal to or greater than 0.8.

Unfortunately, none of these three models provides for selection and assignment on LSE, the optimal process that must be implemented if the MPP standard score is to measure PUE accurately. While assigning on $u$ provides the same set of assignments as does use of the LSEs as assignment variables, selection would be more efficient if accomplished on LSEs rather than on the unique components, the general components, or the unit weighted sum of $u$ and $g$. A simulation approach is probably required to meet all the conditions for an ideal measure of PUE when there are more than two assignment variables corresponding to two jobs.

Brogden's (1959) Table 1 (our Table 2.2) and the equations that yield Table 2.3 values are valid and practical tools for policymakers and research personnel as shown by several examples. We will use the two-stage selection/classification concept, the "selection on *g*" model as the source of the MPP standard scores used in these examples.

In the first example, we stipulate an SR of 0.70 (30 percent of the youth population is rejected). We assume a given test battery where $r$ equals 0.95 for four composites (LSEs) corresponding to four job families; $R$ for the existing four LSEs used as composites is equal to 0.70. The question posed regarding research strategy is: In the development of a new battery, how much would predictive validity (i.e., $R$) have to be increased to provide a PUE equal to that provided by decreasing $r$ from 0.95 to 0.90? The answer obtainable by the application of simple arithmetic to values from Table 2.3 is that $R$ would have to be raised from 0.70 to 0.78.

Impressive savings in recruiting costs could be obtained from raising the SR from 0.70 to 0.80 (i.e., rejecting 20 percent instead of 30 percent). To retain the same estimated PUE provided by the battery (and four composites) for SR equal to 0.70 ($R = 0.70$, $r = 0.95$), $r$ would need to be lowered to 0.85 or $R$ raised to 0.88, or to some combination of improvement in the values of $r$ and $R$ (i.e., increase in R and/or decrease in r) that would yield an MPP standard score of 0.503.

If researchers could identify an effective additional job family and an equally effective associated test composite (i.e., $m$ raised from 4 to 5), the augmented battery could, for a SR of 0.80, and a smaller increase in R or decrease in r provide a value for PUE that equals the PUE of the old battery with the more expensive SR of 0.70, and thus obtain the desired reduction in recruiting costs without a loss in PUE. Raising $m$ from 4 to 5 provides several advantages: (1) the original level of PAE could be retained by decreasing $r$ from 0.95 to 0.89 (instead of decreasing r to 0.88, required for $m = 4$) or (2) increasing $R$ from 0.70 to 0.84 (instead of the 0.88 required for $m = 4$). Increasing $m$ cannot be expected to provide the increases indicated by the "selection on *g*" modification of Brogden's model unless one of Brogden's more important assumptions is met: a joint predictor-criterion space with a dimensionality equal to or greater than $m$ is present. In practice it is difficult to achieve an $m$ greater than three, and probably impossible for $m$ greater than six until major breakthroughs in test research occur. The gains obtainable from using LSEs for each job instead of for job families usually accrue more from the improvement in job clustering than in the increasing of the joint predictor-criterion space dimensionality. Thus, there is nothing inconsistent in expressing caution concerning the use of m greater than 6 for the application of Brogden's model and the recommendation that thirty or forty LSEs be substituted for the existing nine aptitude area composites used in Army classification.

## C. HORST'S CONTRIBUTION TO THE MEASUREMENT OF PAE/PCE

Horst (1954) provides a measure of classification prediction efficiency which he calls an index of differential prediction efficiency. His measure is a psychometric index, an indicator of differential validity, in contrast to absolute or predictive validity. However, Horst's index can be directly linked to MPP, and thus to utility, through its relationship to Brogden's measure of classification efficiency.

Horst (1954) states: "In order to develop a method for selecting that subs· of predictors of specified size which will yield the most accurate predictions of differences between all pairs of criterion measures we must first define mathematically an index of differential prediction efficiency of a test battery" (p. 3). Horst then notes that LSEs can be substituted for the (unobtainable) criterion measures. Horst and others were using this relationship before Brogden published his rigorous proof of the theorem. This issue was discussed in more detail in Chapter 1.

Describing his index in terms of the separate LSEs for each job, Horst states: "The index of the differential prediction efficiency of the battery is taken to be a simple function of the average of the variances for the predicted difference scores for all possible pairs of criterion variables" (p. 3). Thus, Horst's index is equal to the average squared difference between each pair of predicted criterion measures, assuming standard measures for both predictors and criteria and that the predicted criteria are the "least square estimates" (LSEs).

We refer to Horst's differential index as $H_d$, and continue to use the same notation for the covariances of the LSEs (i.e., the matrix of LSE covariances is **C**). Horst states that $H_d$ is equal to a function of the difference between the average diagonal value (or element) and the average off diagonal (or element) of **C**. Horst provides a more precise definition of $H_d$:

$$(H_d) = (\mathrm{tr}\ \mathbf{C}) - \mathbf{1}'\mathbf{C}\mathbf{1}/m \quad ; \qquad (8.3)$$

where tr stands for trace and "tr **C**" stands for the sum of the diagonal elements of **C**, the 1s are column vectors with each element equal to one (summing vectors), and $m$ is equal to the order of **C** (number of jobs).

In other words, $H_d$ is equal to $(m-1)$ times the difference of (1) the average of the diagonal terms and (2) the average of the off diagonal terms of **C**. Thus, the values of **C** given Brogden's assumptions yield a value for $H_d$ of $(m-1)(R)^2(1-r)$. Using Brogden's estimate of PAE and expressing it in terms of $H_d$ (see Appendix 2F) the following relationship holds:

$$\text{PAE} = \left(M_{0m}/\sqrt{m-1}\right)(H_d)^{1/2} \quad . \tag{2.4}$$

Since the terms to the left of $H_d$ remain constant for all computations in which the number of jobs are equal, it can be seen that the use of Brogden's estimate of PAE would yield the same order of merit for any set of alternative batteries as would the use of $H_d$ in a situation in which Brogden's assumptions are met. Thus a direct link is established between MPP, PAE, and $H_d$, and $H_d$ becomes, given Brogden's assumptions, a measure closely related to utility, rather than just a psychometric index.

The simple formula for $H_d$ defined above in terms of **C**, (Equation 2.3) is the final form resulting from a lengthy derivation by Horst. The formulation of $H_d$ most useful for use in one of his two sequential test selection procedures is more closely related to his beginning concept, that is, to the prediction of the $m(m-1)$ non-null difference scores among $m$ LSEs. We find it highly useful to explore further the even more general concept of $H_d$ as a function of the squared correlations of differences between "best" predictors with the corresponding differences among criterion scores.

A factor solution of the correlation matrix among tests (with ones in the diagonals and designated as $\mathbf{R}_t$) can be completely factored so that $\mathbf{F}_t\mathbf{F}_t' = \mathbf{R}_t$. $\mathbf{F}_t$ can be extended (Dwyer, 1937)[9] to the $m(m-1)$ non-null differences among the LSE to provide a Dwyer factor extension solution ($\mathbf{F}_d$); each column element in $\mathbf{F}_d$ is the correlation of a difference variable with the factor represented in the same column in $\mathbf{F}_t$. Each factor is a variable with variance of one that is orthogonal to the other factors. The variable corresponding to each row of both $\mathbf{F}_t$ and $\mathbf{F}_d$ can be thought of as a dependent variable predicted by the vector of regression weights found in that row and applied to the column variables (factors) that serve as the independent variables. Thus, each row of $\mathbf{F}_d$ is a vector of regression weights.

---

9 The factor extension concept is defined and a solution provided in Dwyer's (1937) article; $\mathbf{F}_d$ is an extension of $\mathbf{F}_t$ into the criterion space (in this case specifically to the differences among the criterion variables).

$\mathbf{F}_t$ can be similarly extended to the $m$ criterion variables represented by the rows in the matrix called **F**. This factor solution, **F**, is very important because: it can also directly yield a value for $H_d$; it is easier to use; it is central to the test selection process; and with selected orthogonal rotations it can supply several interesting and useful factor solutions. The derivation and further exploration of **F** can be found in Appendix 2C.

It can be shown that the regression weights of the row variables in **F**, the predicted criterion variables, can provide the means for computing the rows of regression weights for predicting the $m$ $(m-1)$ non-null criterion difference variables. Each difference vector (e.g., representing the difference between the $i^{th}$ and the $j^{th}$ criterion variables) can be computed by subtracting the $i^{th}$ row from the $j^{th}$ row of **F**. When all $m^2$ differences among variables (including the $m$ null variables) are considered in $F_d$, the variance of each column of $\mathbf{F}_d$ is equal to the variance of that column in $\mathbf{F}_d$. Since $H_d$ is defined as the sum of the column variances of $F_d$ divided by 2m, $H_d$ is also equal to the sum of the column variances of $F$. (See Appendix C.)

It is convenient to define $H_d$ in matrix notation. The use of matrix notation will make it obvious that **F** under all orthogonal rotations will yield the same value for $H_d$. Thus, **F** need not be expressed in terms of the triangular factor solution implied by Horst's test selection procedure, but instead any orthogonal rotation of **F** will suffice to yield $H_d$. This matrix formula is as follows:

$$H_d = \text{tr}\,(\mathbf{F} - \mathbf{HF})(\mathbf{F} - \mathbf{HF})' \quad . \tag{2.5}$$

**H** is an operating matrix with all elements equal to $1/m$, and the same number of rows and columns as **F**.

Appendix 2C shows that Equation (2.5) in this general form is algebraically equivalent to Horst's final formula for $H_d$. Thus, and noting once more that $\mathbf{FF}' = \mathbf{C}$, we can use the following relationships:

$$H_d = \text{tr}\,(\mathbf{F} - \mathbf{HF})(\mathbf{F} - \mathbf{HF})' = \text{tr}\,\mathbf{C} - (\mathbf{1}'\mathbf{C}\ \mathbf{1})/m \quad . \tag{2.6}$$

The further relationship of $\mathbf{FF}' = \mathbf{C}$, which follows from the identification of **F** as a Dwyer factor extension solution, equates this **F** with the **F** used to reproduce a covariance matrix **C**; one particular class of **C** matrices are the covariance matrices whose corresponding correlation matrices have values for all its elements that meet Brogden's assumptions.

There is a need for a general formulation of an index analogous to $H_d$ that will provide the maximum flexibility while remaining true to the basic concept of measuring the presence of differential validity in a set of predictors. McLaughlin et al. (1984) pointed out the difficulties in using $H_d$ as an index of classification effectiveness of an operational assignment procedure in which test composites that are not LSEs have prescribed matches with jobs and in which the test composites are used in the assignment process as surrogates for predicted performance with respect to predesignated jobs. This operational constraint prevents the use of LSEs as assignment variables and should be reflected in the estimation of potential operational classification efficiency. The situation of interest is defined in terms of the existing aptitude areas and job families. In other words, the maximum MPP standard score obtainable using an optimal assignment process should be estimated under the restriction that specifies aptitude area composites are (and will continue to be) used as the assignment variables for stipulated jobs.

Horst's index, $H_d$, as noted earlier, can be most generally stated as the sum of the squared correlations between the difference between each pair of criterion scores and the best predictor of each such difference obtainable from the battery. Horst also defined $H_d$ in terms of the criterion scores, **Y**, as the sum of the $m(m-1)$ values equal to $(Y_j - Y_k)^2$, with $j$ and $k$ ranging over all values from 1 to m, and then averaged over the $N$ individuals in the sample. This latter definition holds only because the "best" predictor stipulated by Horst is the difference between the two LSEs associated with the two criteria whose difference is being predicted. Thus, each pair of LSE differences is being correlated with itself and the above simplified formula holds.

A more general formulation of an index of differential validity can be defined; such a general index requires the computation of each correlation coefficient between the criterion pairs and the designated predictors of these pairs. The covariances of each predictor pair and each criterion pair would be summed without squaring in order to preserve the sign of each cross product in the computation of their average value (the differential validity). Several alternative indices of differential validity analogous to $H_d$ are discussed in a following section.

McLaughlin, Rossmeissl, Wise, Brant, and Wang (1984) suggested that it would be interesting to examine a modification of Brogden's assumptions (of equal intercorrelations and equal validities, etc.), with all assumptions fully retained, except that the validities (the $R_i$s) be permitted to vary. The authors give a formula for $H^2$ separated into an alleged Brogden measure and a component of differential validity due to the variation in predictability of the criteria; what they call the "Brogden measure" is defined as $R\sqrt{1-r}$ , but without using the value for $M_{pm}$, and without concern that Brogden's measure is undefined for this situation. The other component of their index involves $m$, the number of jobs, r, the average intercorrelation of the LSEs, and the variance of the $R_i$s across jobs. Computing $H_d$ from **F** with values prescribed by changing the assumption regarding the permissibility of variation among the $R_i$ provides interesting results, although different from those given by McLaughlin et al. (p. 46).

We redefine **F** using Brogden's assumptions, except that the validity for the $i^{th}$ LSE is $R_i$. The standard deviation of the values of $R_i$ will be referred to as $S_R$ and the mean of the $R_i$ denoted as $\bar{R}$. Using this notation the $i^{th}$ row of **F** has a first element of $R_i\sqrt{r}$, has one element equal to $R_i\sqrt{1-r}$ and all other, $m-1$, elements are equal to zero. The non-zero elements to the right of the first column form an $m$ by $m$ diagonal matrix section. The **C** matrix, as reproduced by **FF'**, has diagonal elements of $R_i^2$ and off diagonal elements equal to $(rR_iR_j)$.

When the formula, $H_d = \mathrm{tr}\,(\mathbf{F}-\mathbf{HF})(\mathbf{F}-\mathbf{HF})'$, is used to compute $H_d$ from an **F** containing $\bar{R}$ substituted for each $R_i$, the result is $H_d = (m-1)(\bar{R})^2(1-r)$. The result would be the same, of course, from computing $H_d$ from the values in the matrix **C** using equation 2.3. When the **F** reflecting unequal values for $R_i$ is used to compute $H_d$ the terms in the expression for $H_d$ can readily be separated into a term equal to: (1) $(m\text{-}1)(\bar{R})^2\ (1-r)$, that is, equal to what would be obtained for $H_d$ if $\bar{R}$ had been substituted for $R_i$; (2) a separate term derived entirely from the $g$ factor, and (3) a third term deriving from the $m$ unique factors.

The first of these three terms could be considered a measure of allocation efficiency and labelled $H_{ub}$. The second term, $H_g$, appears to be a measure of hierarchical classification efficiency derived from the $g$ factor, and the third term is the hierarchical classification efficiency derived from all $m$ of the unique factors ($H_{uc}$). Thus, $H_d = H_{ub} + H_g + H_{uc}$. These three components of $H_d$ are as follows:

$$H_{ub} = (m-1)\,(\bar{R})^2\,(1-\mathrm{r}) \quad ; \tag{2.7a}$$

$$H_g = (m)\,r\,(S_R)^2 \quad ; \tag{2.7b}$$

$$H_{uc} = (m-1)\,(1-r)\,(S_R)^2 \quad ; \tag{2.7c}$$

A simplification of the above, which keeps only the contributions of $u$ and $g$ separate, yields the following relationships:

$$H_u = H_{ub} + H_{uc}\ ;\ H_d = H_g + H_u\ ; \tag{2.8a}$$

$$H_d = (m)\ r\ (S_R)^2 + (m-1)\ (1-r)\ (\textstyle\sum_i^m R_i^2)\ \ ; \tag{2.8b}$$

For $r$ equal to one, a special situation in which **F** has only one non-null factor exists (i.e., all the non-zero coefficients of **F** are in the $g$ factor column); thus, all the contribution to classification efficiency is due to hierarchical classification. For the case described immediately above, $H_d = H_g = (m)\ (S_R)^2$. The MPP standard score can also be readily computed for this special case in which there is no contribution to PCE from allocation.

For $S_R$ equal to zero, all of Brogden's assumptions are met in our example, and **F** is still defined as above; $H_d = H_{ub} = (m-1)\ (\overline{R}\ )^2\ (1\text{-}r)$. The tables based on the "selection using $g$" modification of Brogden's model provided in the previous section can be used to obtain an estimated PAE for such an example.

Continuing to use the same notation, we consider four examples in which all of Brogden's assumptions are met, except that the $R_i$ are permitted to be different. These examples are given below; values for $R_i$ (or $\overline{R}$ ), $m$, and $r$ are specified. Seven jobs, $m$ = 7, are stipulated in these examples. It is assumed that a noımally distributed youth population from which 70 percent are selected for further classification is the basis for computing the MPP standard score that would result from the selection and assignment of eacn example. $H_d$ is computed for the first two examples where $r$ is set to one; a value for $r$ will be selected for the third and fourth examples so as to make their $H_d$ values equal to those computed for the first and second examples. These four examples are compared in Table 2.5.

The purpose of providing values of $H_d$ and MPP standard scores for the four examples described above is to demonstrate the discrepancy between computed values for MPP and $H_d$ when the value of the index is based on hierarchical classification effects. For the extreme values of $r$, $r$ = 0 and $r$ = 1, and a moderately large spread of $R_i$ values, as in the first two of our four examples, there is no basis for assuming that $H_d$ is proportional to PCE. We see little justification in using $H_d$ as an estimate of PCE in situations where hierarchical classification effects are a major contributor to PCE.

It will be noticed immediately that, although the first two examples have different values for $\overline{R}$ , both examples have the same value tor $S_R$, and, since $r$ = 1, $H_d = (m)(S_R)^2$. Thus, $H_d$ is the same for the first two examples and a value of $r$ is selected for the third and fourth examples such that all four examples will have the same value for $H_d$.

**Table 2.5. A Comparison of Three Examples Having Equal $H_d$ Values But Unequal MPP Values**

| Example Number | Average Validity of LSEs ($\bar{R}$) | Average Inter-r Among LSEs ($r$) | Standard Deviation of $R_i$ | Value for ith layer (all layers have equal $N$s) ($R_i$) | $H_d$ [a] | MPP Standard Score |
|---|---|---|---|---|---|---|
| 1 | 0.5 | 1.0 | 0.1 | 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65 | 0.7 | 0.31 |
| 2 | 0.4 | 1.0 | 0.1 | 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55 | 0.7 | 0.27 |
| 3 | 0.5 | 0.533 | Ø | 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5 | 0.7 | 0.46 |
| 4 | 0.4 | 0.271 | Ø | 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4 | 0.7 | 0.46 |

NOTE:

a $H_d$ is entirely due to $H_g$ in examples 1 and 2, and entirely due to $H_{ij}$ in examples 3 and 4. If one assumes an $H_d$ of 0.7, for m = 7, to mean MPP is equal to 0.46, and seeks to generalize this result to examples 1 and 2, the actual MPP values for these examples would be greatly overestimated by $H_d$ (assuming proportionality of $H_d$ and MPP).

The easy computation of a value of MPP for the first two examples requires the use of a normal distribution. This makes it most convenient to base all examples on the youth population where the required normal distribution exists by definition. To make the examples resemble the real world, 30 percent will be non-selected. In Table 2.5[10] the value of MPP was designated as $(MPP)_y$ when MPP is computed in forms of scores that have a zero mean in the youth population, and as $(MPP)_w$ when MPP was computed on scores whose mean equals zero in the selected enlistee population. The values of $(MPP)_w$ provided the closer relationship to $H_d$.

For the first two examples, the values for $(MPP)_w$ were obtained by subtracting the mean of a randomly assigned group from the value of $(MPP)_y$. The latter were computed by using a normal curve table and a simple formula for computing MPP in a placement model. The means of each successive segment of the normal curve, first for those above

10 The values for $(MPP)_w$ were not included in Table 2.5; $(MPP)_y$ is shown labeled as MPP.

the 90th percentile, then for those falling between the 80th and the 90th percentiles, etc., are computed. The highest mean will be multiplied by the highest $R_i$, the next highest mean by the next highest $R_i$ until the 30th percentile (the rejection point) is reached. The sum of these seven products provides a value for $(MPP)_y$, which can be converted to $(MPP)_w$ as discussed above. MPP values and the $H_d$ values are provided for each of the four examples identified in Table 2.5.

Table 2.5 shows that as $r$ approaches 1, and $S_R$ is moderately large, the contribution of any hierarchical classification effect to MPP is substantially less than its contribution to $H_d$. The presence of hierarchical classification effects inflates $H_d$ while having much less effect on MPP standard scores. As a corollary to this statement, the link between $H_d$ and MPP estimated from the use of Brogden's model becomes tenuous when a substantial amount of hierarchical classification effect is present and $r$ does not approximate zero. The substitution of a mean $R_i$ (when all $R_i$ do not equal $\overline{R}$ ), in order to use the Brogden tables directly, is made doubtful by the results shown in Tables 2.5 and 2.6. One can be confident of the proportional relationship between $H_d$ and a squared MPP only when $H_d = H_{ub}$.

When $r$ is less than one, a contribution of hierarchical classification effect can also be provided by the unique factors, not just by the general factor, as was true for examples 1 and 2. When $r$ equals 0.8, $\overline{R}$ equals 0.5, and $S_r^2$ equals 0.1, the contribution of $H_g$ to $H_d$ is 63 percent of its total value, while $H_{ub}$ and $H_{uc}$ contribute, respectively, 34 and 13 percent. This large component of $H_g$ in $H_d$, when $r$ is high, is significant because the results of Table 2.5 indicate that $H_g$ does not share the close relationship of $H_{ub}$ to MPP. We suspect that $H_{uc}$ lies between $H_g$ and $H_{ub}$ with respect to their relationships to MPP. Table 2.6 provides a shredding out of $H_d$ into its components for several modifications of examples 1 and 2 described above. The modification of the examples is only with respect to the value assigned to $r$; the examples are left unmodified for the last two rows of Table 2.6 (i.e., $r$ = 1.0).

In summary, $H_d$ has a direct link to MPP, and thus to utility measures, only when the classification effectiveness of a battery is entirely due to its allocation effectiveness. Operationally this condition would exist if existing aptitude area scores were used in the assignment process, rather than giving them a variance proportional to the variance of the LSEs (the squared multiple correlation coefficients), and differential cut scores (minimum prerequisites) are not used.

Table 2.6. Three Components of $H_d$

| Example Number | Average Validity of LSEs ($\bar{R}$) | Average Inter-r Among LSEs ($r$) | Range of $R_i$ | $H_g$ | $H_{ub}$ | $H_{uc}$ | $H_d$[a] |
|---|---|---|---|---|---|---|---|
| 1 | 0.5 | 0.80 | 0.35 to 0.65 | 0.56 | 0.30 | 0.12 | 0.98 |
| 2 | 0.4 | 0.80 | 0.25 to 0.55 | 0.56 | 0.19 | 0.12 | 0.87 |
| 3 | 0.5 | 0.90 | 0.35 to 0.65 | 0.63 | 0.15 | 0.06 | 0.84 |
| 4 | 0.4 | 0.90 | 0.25 to 0.55 | 0.63 | 0.10 | 0.06 | 0.79 |
| 5 | 0.5 | 0.95 | 0.35 to 0.65 | 0.66 | 0.08 | 0.03 | 0.77 |
| 6 | 0.4 | 0.95 | 0.25 to 0.55 | 0.66 | 0.05 | 0.03 | 0.74 |
| 7 | 0.5 | 1.0 | 0.35 to 0.65 | 0.70 | 0 | 0 | 0.70 |
| 8 | 0.4 | 1.0 | 0.25 to 0.55 | 0.70 | 0 | 0 | 0.70 |

NOTE:

[a] The sum of $H_g$ and $H_{uc}$ remains almost constant for values of r between 0.8 and 0.95; the increase in $H_d$ over this range is primarily due to $H_{ub}$, the component of $H_d$ we believe best reflects MPP.

The linkage of $H_d$ with MPP appears robust enough to justify the use of $H_d$ for test selection purposes, but not necessarily robust enough for use as a measure of MPP standard scores (as the first step in estimating utility). Caution should be exercised in the use of $H_d$ as a measure of PCE. This is especially true when used under conditions that maximize hierarchical classification effects, when the intercorrelation of LSEs (or composites) are high, or when comparisons are being made across sets of test composites that imply different values for "$m$" (dimensionality of the joint predictor-criterion space) across composite sets. All three of these conditions that counter-indicate the use of $H_d$ as a measure of PCE occurred in the Project A analysis cited above (McLaughlin et al., 1984).

## D. IMPACT OF BROGDEN'S AND HORST'S CONTRIBUTIONS ON THE SUBSEQUENT CLASSIFICATION LITERATURE

It is unfortunate that authoritative reviews of the classification literature have discussed Brogden's 1951 and 1959 articles in such a way as to lead their readers to misinterpret some of his terminology. Part of the difficulty must be attributed to Brogden's attractive tables that gave the erroneous impression of being self contained. Authors who used Brogden's tables in journal articles or text books did not provide a definition of the term "predictor" used in key column headings of these tables, although Brogden had defined his "predictors" as necessarily LSEs in the text of his articles (most clearly in a footnote in his 1951 article). In his 1959 article he discussed at length why tests or test composites must not be substituted for LSEs.

In Brogden's 1951 and 1959 models for the measurement of classification efficiency, he was, in effect, defining PAE within his somewhat limiting assumptions. His tabled values of PAE in the form of MPP standard scores divided by $R\sqrt{1-r}$, as defined in the previous section, could be obtained by entering with the selection ratio and the number of jobs for which "predictors" with a unique component extending into the criterion space are available. Values for $r$, the common intercorrelation coefficient among the "predictors," and for $R$, the common validity coefficient for each job predictor against each job criterion, must be specified to produce a value for PAE from Brogden's table. Brogden's precise definition of these predictors as LSEs based on the total battery is essential to the meaningfulness of these tables.

As previously noted, the PAE indicated by Brogden's tables is difficult to achieve in practice. This potential is within reach only when LSEs are used as the predictors; the equivalent of a LP algorithm is used to make both rejections and assignments; the assumption of a dimensionality in the predictor-criterion space greater or equal to one more than the number of jobs is met; and all LSEs are equally valid against their corresponding (target) job. If the second assumption is not met, an underestimate of PCE results. Conversely, the next to last assumption, one that is seldom met for more than 3 or 4 jobs, will cause the entries from Brogden's Table 1 (1959) to overestimate PAE. The use of correlations among test composites instead of LSEs as the argument in the multiplier $\sqrt{1-r}$ will usually cause a moderate to severe overestimate of PAE.

Hunter and Schmidt (1982) state that Brogden's "classic study" used a function with dependent variables which included the validities and intercorrelations of "estimates of job performance." Thus far this is correctly stated. Unfortunately, they added the following two sentences a few lines below: "Brogden had in mind the case in which job performance is predicted using regression equations derived on a common battery of tests. However, the model also holds when a different test is used to predict performance on each job." (p. 259.) This last sentence is clearly wrong and could lead the reader to believe erroneously that validities and intercorrelations of Army aptitude area composites may be substituted to obtain estimates of MPP standard scores (PAE) in using Brogden's tables.

Hunter (1986) first states, "Brogden (1959) quantified the gains that would arise from optimal classification and showed that gain depends strongly on the size of the correlation between the aptitude composites tailored to different jobs." (p. 356.) Hunter correctly notes the difficulty of keeping the intercorrelations among test composites low (and even more difficult, we add, if the composites are LSEs). Hunter then states that: "The only way to keep these correlations in the 0.80s or low 0.90s is to restrict the number of tests in each composite and to artificially make the composites as close to non-overlapping as possible." (p. 356.) Conversely, Brogden was very emphatic that it was *only the intercorrelations of LSEs*, not of other test composites, that have a proven relationship to PAE. Brogden also provided definite proof that the removal of tests from LSEs to avoid overlapping of tests, or for almost any other reason, will reduce PAE. The only situation in which test removal can be effected without resulting in a reduction of PAE, is when the regression weights (Betas) for the test to be removed are equal across all jobs. If this latter condition holds, the reconstitution of the test composite by removing one or more tests is, of course, called for, and their removal can scarcely be called artificial.

Cascio (1987b) also provides one of Brogden's tables as an illustration of the potential effects of classification. He presents a table titled "Mean Standard Criterion Score of Persons Placed on Two Jobs By Placement Or Classification Strategies," adapted from Brogden (1951). The column headings representing intercorrelations of two LSEs are labeled, after Brogden, as "Two Predictors Whose Intercorrelation Is." Presenting this table without an explicit explanation that predictors cannot be just any test or test composite, but must be LSEs, has a high potential for misleading the reader.

Cascio cited Anastasi (1982) for the rule that, "... a classification battery requires a separate regression equation for each criterion." (p. 338.) Unfortunately, he adds, "The particular combination of predictors employed out of the total battery, as well as the specific weight given each predictor, varies with each criterion" (p. 338). In his discussion of Brogden's 1951 model, Cascio writes of the possible use of "separate predictors (or regression equations) for each job." (p. 338.) A reader may erroneously interpret his words to mean that an option is permissible: either to use data pertaining to predictors or regression equations in entering Brogden's table. Rather, the reader should be led to understand the necessity for the LSEs to be based on the full battery. Particular combinations of tests selected out of the total battery are not permissible for the computation of intercorrelations among predictors (LSEs) when these intercorrelations are to be used as the "$r$" in Brogden's model.

Cascio's description of "the multiple regression model based on differential performance" (p. 340) is correct for the two-job case but does not generalize to the assignment of individuals to three or more jobs. Cascio does not state that the usefulness of equations predicting difference scores would generalize to a situation with more than two jobs, but he does not inform the reader that it would not. The use of a regression equation that predicts the difference between the LSEs for each of two criterion variables can be used to determine an individual's "relative fitness for job A over job B," but the operational usefulness of three such equations when there are three jobs, or six such equations when there are four jobs, is almost nil.[11]

A more general approach to the assigning of individuals to jobs, one that works well with two or more jobs, is provided by using the LSEs that predict each criterion as the basis for making assignments. In the two-job case the predicted performance scores for each job (LSEs) are computed for each individual and an appropriate constant added to the job with the larger quota. Each individual can then be assigned to the job corresponding to his highest score. Exactly the same assignment decisions would result as those made by using the regression equation predicting the difference between the two criterion scores. However, this simple approach generalizes to the assignment to three or more jobs. In the general case, an appropriate constant is added to each predicted performance score to reflect the desired quotas when the individual is assigned to his highest adjusted score. Since this is the multidimensional screening (MDS) procedure, a provision for rejecting either a prescribed number of applicants, or all applicants with less than the required predicted performance score, can be conveniently utilized.

Cronbach and Gleser (1965) distinguish between "general" and "differential" predictors, implying Brogden's (1951) predictors were something quite different from the

---

[11] We are here distinguishing between the usefulness of predictor-criterion differences in the operational assignment problem as contrasted with the test selection process.

LSEs that maximize the prediction of each criterion variable using full regression equations that include every test in the battery. Brogden did not specify the number or nature of the tests in the battery, only the number of jobs to be predicted with separate LSEs. Cronbach and Gleser doubted that differential predictors of the type they believed are essential to Brogden's model would be as valid as a general predictor. They write: "Brogden's example, one should note, assumes that differential predictors have the same level of validity as the general predictor, which is not true for differential batteries so far developed" (p. 112). Quite the contrary, the average validity of a test composite that is valid across a number of jobs (a general predictor) could not exceed (or equal, unless the predictor criterion space was unidimensional) the average of the validities of the separate LSEs for each job. It does not appear that those LSEs that have the maximum validity obtainable from the battery for each job are the differential predictors referred to by Cronbach and Gleser. Brogden, however, (1951, 1959) had no other predictors.

Even when each job is considered separately in the selection procedure, the LSE predictors appropriate for use in the univariate (one job) case are the same as the predictors used for that same job in the multivariate (two or more jobs) case. The predictors in Brogden's model are always the LSEs based on the total battery and are the same for a given job, regardless of the number of jobs being considered in the classification and/or assignment process.

Horst's classification efficiency index, $H_d$, is generally neglected in the literature on utility analysis. Cronbach and Gleser (1965) and McLaughlin et al. (1984) are two notable exceptions. The latter creatively used an extension of the Horst index to compare the classification efficiency of several alternative sets of test composites drawn from the ASVAB. This suggested modification is discussed later in this section. Cronbach and Gleser provided what is probably the best known review of the Horst index. We now consider the accuracy of this review.

Cronbach and Gleser (1965) writing about Horst's differential index state that: "certainly his procedure does not provide the ideal battery for fixed-treatment classification." (p. 118) They add that "The function used to define efficiency does not correspond clearly to any common type of decision problem, and it is demonstrably not the correct function for the fixed treatment example to which Horst applies the method" (p. 119). Neither Brogden nor Horst related their measures of classification efficiency to each other; exercising caution about a method which did not purport to provide results in terms directly relatable to utility was appropriate. However, since we now know, as demonstrated earlier in this chapter, that $H_d$ is directly proportional to the square of PAE when Brogden's assumptions are met and the number of jobs is held constant, it can be said that $H_d$ has much more merit as a surrogate for utility in a test selection procedure than was perceived by Cronbach and Gleser nearly twenty years earlier.

Cronbach and Gleser (1965) also make the objection to the usefulness of $H_d$: "The Horst solution, moreover, makes no adequate provision for a reject group who receive no courses, job assignments, etc. Thus his analysis would apply only when all individuals tested are to be utilized." (p. 118) However, we claim that in the selection of tests for a battery, it is essential to consider classification efficiency whenever both selection and classification is accomplished, in either two separate stages or in a single simultaneous stage. When selection is part of a personnel utilization procedure that also includes assignment to two or more jobs, the use of $H_d$ (or some more or equally effective PCE index) is necessary but not sufficient.

In a two stage selection/classification multi-job model, the selection of tests to be used in the selection stage could appropriately be accomplished using Horst's absolute validity index, $H_a$, which optimixes selection efficiency, and a separate classification battery identified using $H_d$ (Horst's differential index). There would be no need to make a selection decision in the classification stage nor a classification decision in the selection stage of such a model.

When, either: (1) further selection is to be accomplished in the classification stage (i.e., use of minimum entry requirements for particular courses or programs), or (2) selection and classification is to be accomplished simultaneously, as in the MDS process, provision for the efficient identification of rejects is essential. It also is essential that classification efficiency not be reduced in the quest for PSE. Fortunately, classification efficiency need not be reduced to achieve greater selection efficiency. This is demonstrated in the notional example described below.

In the example, we use the MDS procedure and a battery in which tests were selected to maximize $H_d$. This procedure, also discussed in Chapter 7, assures that no rejected individual can have a higher predicted performance score for a given job than any individual retained and assigned to that job using an optimal assignment algorithm. This cannot generally be assured in two stage selection/classification procedures.

In this example, an LP program used with tentative quotas proportional to the actual job quotas is used to make a trial optimal assignment of every applicant to a job. The tentative quotas are inflated proportionately to allow for rejects. Assignments are made by adding a job (column) constant to the predicted performance scores (LSEs) of each individual to obtain an adjusted score; each person is then tentatively assigned to his highest adjusted score. With the proper selection of job constants, quotas will be met and the MPP standard score maximized. Next, each person is rank ordered on his adjusted predicted performance score (the score corresponding to the job to which he is tentatively assigned), and a count from the top scorer down made to retain a sufficient number of individuals to meet the actual quotas. The point at which the quotas are met is the cutting score below which individuals are to be rejected. Thus no individual in the rejected group could be used to replace an accepted and assigned individual without lowering the MPP standard score. The test composites best for selection are also best for classification, since the LSE is best for both purposes. Thus, no further increase in selection efficiency can be accomplished using any other composite from this battery, one designed to maximize $H_d$, and, hopefully, PCE.

The addition of one or more tests selected to maximize $H_a$ can provide for increasing selection efficiency. If, for this augmented battery, the test composites best for both procedures (i.e., separate LSEs based on the full battery for each job) are used in the MDS procedure, no loss in classification efficiency will result while the PUE will increase (as a result of greater selection efficiency).

We will hypothesize the addition of a measure of general mental ability to a battery that was selected to maximize $H_d$. We will further assume that the addition of this test will appreciably increase the $H_a$ of this augmented battery while adding nothing to the battery's $H_d$. This would be true if general mental ability was equally valid for all jobs and was not already measured by some combination of the tests initially selected for inclusion in the hypothetical battery.

Under the above assumptions a recomputation of the LSEs for each job and a fresh application of the MDS procedure would leave every one tentatively assigned as before but rank ordered differently on the adjusted LSEs, thus providing a different set of rejectees. Even the rejectees would have the same tentative assignments as before, but some individuals rejected through use of the initial, less valid, LSEs would now be identified as appropriate for acceptance. In this example it would appear that both the $H_d$ and the $H_a$ indexes were effective for the purposes intended by Horst. His $H_a$ and $H_d$ indexes were surely intended to be supplementary approaches, and $H_d$ should not be expected to accomplish the role of $H_a$, or vice versa.

The McLaughlin et al. (1984) study is the only formal technical report issued during the first six years of "Project A." In this study, the investigators used a Horst type measure of differential validity as an estimate of the potential classification efficiency obtainable from specified sets of aptitude areas. We first describe very briefly their classification methodology and results and then provide a critical evaluation of their modification of the Horst index as a means of comparing alternative sets of Army Aptitude Areas (AAs) in terms of potential classification efficiency. We do not discuss their overall study results.

The alternative AA sets compared in this study include a set of one composite (i.e., as if a current version of the AGCT were to be used in place of its successor, the ACB), sets of 2, 3, and 4 AAs, the then current set of 9 operational AAs, and finally a proposed revision of this set of 9 operational AAs. Despite our reservations concerning the means of evaluating potential classification efficiency, the results of this report are not only the best available but also have significant and fascinating implications.

The Project A study described in this report provided for the collection of both operational and experimental data on over 60,000 soldiers and 98 jobs (MOS). Only the existing ASVAB tests were considered in research to determine the advisability of reconstituting the operational AAs and restructuring Army job families.

McLaughlin et al. (1984) used an average of the Horst differential efficiency index ($H_d$), designated by them as $H^2$, and a creative extension of the concept of $H^2$, designated as $M^2$, to measure the potential classification efficiency of the alternative AAs. The ratio of ($M/H$) was proposed by the authors as an estimate of the percentage of total differential validity that could result from the optimal use of aptitude areas as compared to the optimal utilization of the ASVAB (i.e., the use of 98 LSEs) to assign soldiers to the 98 jobs using an assignment algorithm that maximizes the LSEs of assigned personnel. They refer to this percentage as "relative efficiency," and say that it assesses "the extent to which the composites capture the differential validity possessed by the ASVAB." (p. 49.)

The computational procedures devised by the authors included several desirable refinements in algorithms used for $H^2$ and $M^2$. For example, alternatives were provided for both algorithms in which the number of soldiers assigned to each job are taken into account. Also, the LSEs for performance on each of the 98 Army jobs are obtained using the ridge equation method to reduce shrinkage of validity of these best weighted equations in future samples (Draper and VanNostrand, 1979). Appropriately in the computation of $M^2$, the same estimates of performance differences are used across the different batteries (i.e., the different sets of AAs). These added computational features make the comparison of $M^2$ values more meaningful across sets of AAs than if an approach similar to that used in Horst's (1954) examples had been utilized.

As described in the previous section, $H_d$ is the sum of the squared correlation coefficients between two differences associated with each pair of jobs. One of these differences (the criterion difference) is between either the actual performance measures or the predicted performance measures (both yield the same result), and the other is the predictor difference, the designated predictor of the criterion difference. Horst prescribed using LSEs as the predictors in his formulation of $H_d$. The Project A authors define the "predictors" as LSEs based on the two AAs corresponding to each criterion pair. The only justification provided for the use of these particular predictors is that: "Each MOS is associated with a single composite, so the comparison of expected performance between two MOS is associated with a pair of composites..." (p. 47). The method has some intuitive attractiveness as being analogous to the use of LSEs based on the total battery as predictors when computing $H_d$. But it is clear that neither the LSEs prescribed by Horst nor the one defined by McLaughlin et al. define the potential for classification efficiency obtainable from an operational assignment procedure that uses AAs as surrogates for predicted performance in a predetermined job family.

The authors reported the "relative efficiency" of the composite set comprised of 98 LSEs (i.e., one per job in lieu of AAs, and measured in terms of $H^2$), as 100 percent (by definition). The current 9 AAs has a "relative efficiency" of 64 percent and a single, AGCT type, composite has a relative efficiency of 43 percent, where the more traditional formulae for $H^2$ and $M^2$, are used (i.e., job samples are not weighted by their size). Additional results are provided in Table 2.7.

The revised set of 9 AAs as recommended in the McLaughlin et al. report, show an 18 percent reduction in the gain of $M^2$ provided by the 9 operational AAs over the single AGCT type composite (again using the unweighted formula). It is noteworthy that the authors considered such a reduction in differential validity an acceptable price to pay for an increase in predictive validity.

**Table 2.7. Differential Validity Indices for Alternative Sets of Test Composites**

| Composite Sets | Differential Index ( $H$ or $M$ )[a] | |
|---|---|---|
| | Traditional Index (unweighted by Job Density) | Index Modified to Reflect Job Density |
| 98 LSEs | 0.314 | 0.214 |
| Current 9 Aptitude Areas | 0.202 | 0.146 |
| Revised 9 Aptitude Areas | 0.190 | 0.142 |
| 4 Composite Set | 0.160 | 0.125 |
| 3 Composite Set | 0.154 | 0.120 |
| 2 Composite Set | 0.150 | 0.125 |
| 1 Composite Set | 0.136 | 0.106 |

Source: Adapted from McLaughlin et al. (1984), pp. 50-51.

NOTE:

[a] $H$ is used for the LSEs and $M$ for all other composites; $H$ is the square root of the mean value of Horst's index of differential validity, thus $H^2 = (H_d)/m$; $M$ lacks a precise relationship to $H$ (see text for description of $M$), but McLaughlin et al. appear to believe that $H$ and $M$ are comparable.

The unmodified Horst method for computing $H_d$ for each of the sets of AAs calls for considering each composite (AA) as a test in a battery of tests, and each set of AAs as equivalent to a battery; the existence of the ASVAB as the source of the test composites in each AA set is immaterial to the computation of $H_d$. Thus a set of $n$ AAs is psychometrically equivalent to a battery of $n$ tests and $H_d$ could be computed for each such battery in the same manner in which the investigators computed an average $H_d$ (i.e., $H^2$) for the battery defined as the set of ASVAB tests. In the discussion that follows we refer to a hypothetical $H_d$ computed for batteries made up of from one to nine AAs as well as the $H_d$ based on the ASVAB.

As noted above, the criterion differences against which the predictor differences are correlated would differ across batteries, making a comparison of values of $H_d$ less meaningful than when the McLaughlin et al. approach (using the same criterion difference score across all batteries) is used.[12] Also, the $H_d$ for each battery would use LSEs, based on the full set of variables in the battery, as the predictors whose differences are correlated with the criterion differences. Since the Army utilizes specified AAs as the assignment variables, rather than these LSEs, an algorithm which uses less valid predictors than the LSEs seems intuitively desirable.

We see no justification for the use of the differences among the two-variable LSEs as the "predictor" differences in the authors' algorithm for $M^2$. We believe the actual assignment variables should be used to compute any index purporting to be an estimate of classification efficiency--the efficiency that could be obtained from the use of specific sets of either LSEs or specified test composites, whichever is to be used in the assignment process.

We propose the use of two alternative procedures for computing a modification of the $M^2$ index. The first we will call $M^*1$ and the second $M^*2$. Since the McLaughlin et al. study assumes that the assignment variables are not in Army standard score form, as are the existing AAs, but instead have standard deviations proportional to their validities, we will, for our first recommended modification, propose the use of predictor variables in this same form. The second alternative modification of $M^2$ assumes that all AAs are in Army standard score form (i.e., all have the same standard deviations across all jobs with no capability either to capitalize on validity differences or to disrupt quality distribution plans). For both $M^*1$ and $M^*2$ we will sum differential cross products considering the signs of all scores. This is in contrast to $M^2$ which is not sensitive to signs since the $m(m-1)$ difference scores for the $i^{th}$ individual are squared before being further used in the algorithm.

In our first alternative modification of $M^2$, we would substitute AA scores, in standard score form, multiplied by their validity coefficients (weights) for the two-variable LSEs used in the McLaughlin et al. algorithm for $M^2$. In our algorithm modification the weighted AA scores corresponding to the $k^{th}$ job would be subtracted from the weighted AA scores corresponding to the $j^{th}$ job (and not vice versa). These predictor differences would be correlated with the criterion score for the $k^{th}$ job subtracted from the criterion score for the $j^{th}$ job. This means that the correlation of a predictor pair with a criterion pair

---

12 $H_d$ is actually a covariance across the differences between pairs of predictor variables and the corresponding differences between pairs of criterion variables: this concept is described in more detail later in the text of this chapter, and again in Appendix 2.

could be negative when the difference between validities for the $j^{th}$ and $k^{th}$ jobs, with respect to their corresponding AAs, yields an opposite sign than the differences between the LSEs based on the entire ASVAB. This is an appropriate result and contrasts with the effect of the McLaughlin et al. algorithm that, in effect, fits error by forcing all signs for the predictor differences to agree with the signs of the criterion differences. The modified algorithms for both $M^*1$ and $M^*2$ will be more precisely defined and discussed in Appendix 2E; our index, one which has the flexibility of incorporating the features of either $M^*1$ or $M^*2$, is identified in the appendix as $H_p$.

For $m$ jobs, each of the $N$ individuals has $m^2$ pairings of predictor difference scores with criterion difference scores. Thus there are $N(m)^2$ pairs of difference scores contributing to the final value of the differential validity coefficient, $H_d$, $H$, or $M$. Each of the pair of scores for producing the predictor differences are potentially different for each individual. Thus there are $m(m-1)$ distinct pairs for each individual, even though the number of separate AA scores for each individual is limited to a number running from 1 to 9, depending on the AA set being evaluated. Similarly the $i^{th}$ individual has $m(m-1)$ potentially separate differences between his criterion scores. With 98 jobs and over 60,000 soldiers in the data set, it is unlikely that recomputation of $M^2$ will be made.

When the number of AAs in the set equals one (the AGCT situation), our proposed modification of the $M^2$ algorithm, if applied to the same data, would yield the same results as would the algorithm of McLaughlin et al. For the set with two AAs (the two test battery), McLaughlin et al.'s algorithm for $M^2$ uses the same predictor differences as would be used for predictor differences in the direct computation of $H_d$ on that two test battery. However, the $M^2$ and $H_d$ algorithms use different values for the criterion differences; the $H_d$ algorithm uses the criterion variables computed from the two test battery while the $M^2$ algorithm uses the criterion variables from the ASVAB test battery. Thus as noted above, the expected values of the differential validity indices, for each set containing two to nine AAs, would be highest for the $H_d$ computed on the ASVAB battery, next highest for the $H_d$ computed on the particular battery, next highest for $M^2$, and lowest for our proposed modifications of $M^2$. As we consider sets with a progressively larger number of AAs, going from three to nine, these four indices have a larger spread, but retain the same rank order.

Our second modification of the $M^2$ algorithm is intended to reflect the classification efficiency obtainable from each set of AAs that have been converted to Army standard score form, and remain unweighted by job validity or value weights, when used in an LP algorithm to assign individuals to jobs. Thus $M^*2$ is like $M^*1$ except that the weights equal to the validity coefficients used in the computation of predictor differences for $M^*1$ are set to one for $M^*2$. Using this modification of the $M^2$ algorithm, the value for $M^*2$ would always be zero for the single composite set and would probably be at least 50 percent smaller for the sets containing from two to nine composites. This modification of the $M^2$ algorithm estimates PAE. Stated differently, $M^*2$ estimates that part of PCE that is due to allocation effects (i.e., PCE with no hierarchical classification effects).

In summary, our first modification of the $M^2$ algorithm would provide a more justifiable estimate of PCE (when composites weighted by job validity are to be used in the assignment process) than does the McLaughlin et al. algorithm for $M^2$. Both algorithms yield the same values for the single composite set. Since the numerator for the "relative efficiency" index remains appropriate for use with either algorithm, our first modification of $M^2$ still yields an efficiency of 43 percent when a single composite is used to make assignments. However, using our modified algorithm considerably lowers the relative efficiency of all sets containing from two to nine composites. The existing and proposed nine AA sets are not nearly as efficient, for classification purposes, as the results provided by McLaughlin et al. indicate.

Our second modification of the $M^2$ algorithm provides a reasonable estimate of the allocation efficiency present in a set of AAs when prescribed AAs are to be used, in unweighted standard score form, as the only estimates of performance on specified jobs, and used in LP algorithms to assign men to jobs so as to maximize the MPP standard score. Such a measure of allocation efficiency lies somewhere between OAE and PAE, not being a measure of the battery's potential but measuring a capacity for operational effectiveness not realized operationally unless an optimal assignment algorithm is used. We know that $H_d$ is proportional to the square of PAE under certain assumptions that includes the absence of a hierarchical classification effect. The existence of a similar linear relationship between PAE and our second modification of $M^2$ seems reasonable. A close linear relationship between PCE and our first modification of $M^2$ seems less likely. Any useful relationship between PCE and $M^2$ seems even less likely.

## E. RULE-OF-THUMB MEASURES OF CLASSIFICATION EFFICIENCY

The most accurate measures of either potential or operational classification efficiency (PCE or OCE) of batteries or sets of test composites are complex to visualize and expensive to realize. Less expensive rule-of-thumb measures that approximated either PCE or OCE would be highly desirable. We describe several candidate heuristics for consideration. The determination of MPP standard scores by simulation or numerical solutions of integrals are expensive procedures that less expensive rule of thumb heuristics seek to approximate. Versions of $M^2$ discussed above are neither sufficiently accurate nor sufficiently inexpensive to be considered as a practical substitute for a simulation approach.

We consider ten rule-of-thumb measures that have been used in previous research to estimate PCE; about half of these rules are at best ineffective, sometimes doing more harm than good. These rules, R#1 to R#10, are summarized in Table 2.8.

**Table 2.8. Figures of Merit Sometimes Used as Measures of Classification Efficiency**

| Rule Number | Figure of Merit [a] | More Appropriate For: OCE | PCE | Accuracy Rating |
|---|---|---|---|---|
| 1 | Composite (or test) intercorrelations | √ | | Low |
| 2 | Predicated performance intercorrelations | | √ | Medium |
| 3 | $R\sqrt{1-r}$ | | √ | High; still needs multiplier reflecting rule #10 |
| 4 | Predicted validity | for composites | for LSEs | Low |
| 5 | $H_d$ | | √ | Medium to High; use with rules #4 and #10 |
| 6 | Comparison of diagonals of $\mathbf{V}_a$ with other row elements | √ | | Very Low |
| 7 | Comparison of diagonals of $\mathbf{V}_a$ with other column elements | √ | | Medium to High for OCE; use with rule #10 |
| 8 | Column variance of **V** | | √ | Medium; a rough estimate of $H_d$ |
| 9 | Dimensionality of either predictor or criterion space | | √ | Low, except at upper bound for rule #10 |
| 10 | Dimensionality of joint predictor-criterion space | | √ | Probably Medium to High if individual factor contributions are considered (but imperfectly understood) |

[a] See text for description of notation.

In defining these "rules" we refer to the matrix of intercorrelations of tests as "$\mathbf{R}_t$" and the intercorrelations of composites (or AAs) as "$\mathbf{R}_a$." The matrix of test validities will be called "$\mathbf{V}$" and the composite validities called "$\mathbf{V}_a$." These matrices have rows corresponding to jobs and columns corresponding to predictors. The diagonal elements of $\mathbf{V}_a$ provide the validities of each composite for its corresponding job (or job family).

Rule #1 implies that a set of composites with lower intercorrelations always will provide a higher OCE than will an alternative set of composites having higher intercorrelations. The index corresponding to this rule is the average of the off diagonal elements of $\mathbf{R}_a$ or $\mathbf{R}_t$. Since this estimate ignores the configurations of validity vectors and predictive validities across jobs, a *set of composites refined by recourse to this rule, may,* particularly with respect to $\mathbf{R}_a$ have its PCE reduced.

Rule #2 uses "$r$" as a measure of classification efficiency and implies the desirability of minimizing $r$. While this is usually good advice and is more useful than rule #1 for evaluating alternative test batteries in terms of PCE, this rule is not relevant to the estimation of the OCE of alternative sets of composites. The value of $r$, as the average intercorrelation of the predicted performance measures, reflects the configuration of validities as well as the intercorrelations of the tests and is thus a useful estimate of the PCE of a battery, especially if selection is not also to be accomplished with the battery.

Rule #3, a very useful rule-of-thumb, assumes that a figure of merit equal to $\overline{R}\sqrt{1-r}$ is closely proportional to the PCE of a test battery. This function was discussed earlier in this chapter. $\overline{R}$ is the average multiple correlation coefficient between all the tests in the battery and each job, and $r$ is the average intercorrelation among predicted performance measures (LSEs). The formulae for $\overline{R}$ and $r$ are based entirely on $\mathbf{R}_t$ and $\mathbf{V}$. Specifically, $r$ is equal to $(1/m)\mathbf{1}'\mathbf{S}\mathbf{V}(\mathbf{R}_a^{-1})\mathbf{V}'\mathbf{S}\mathbf{1}$ where $\mathbf{S}^2$ is a diagonal matrix whose non-zero elements are the diagonals of $\mathbf{V}(\mathbf{R}_a^{-1})\mathbf{V}'$; $\overline{R}$ is equal to the trace of $\mathbf{S}$ divided by the number of jobs ($m$). $\mathbf{R}_a$ and $\mathbf{V}_a$ can be used in computing this figure of merit only if substituted into the above formulae for $\mathbf{R}_t$ and $\mathbf{V}$, respectively; this rule-of-thumb measure is not useful in determining the OCE of existing or proposed AAs, since the value of this measure is based on the assumption that LSEs will be used as assignment variables.

Rule #4 uses several different representations of predictive validity as heuristics, including: (1) the average multiple correlation coefficient, $\overline{R}$ ; (2) the sum of squared multiple correlation coefficients, $H_a$; and (3) the average validity of AAs. Validities are computed either between LSEs or test composites and their corresponding job criteria, and then averaged across all jobs. This rule is based on the assumption that PCE is higher whenever predictive validity is higher, a very erroneous assumption. It is obvious that $\overline{R}$ supplements R#2 and $H_a$ supplements R#5. The average of the diagonals of $V_a$ would be similarly supplemented by our modifications of $M^2$ (i.e., $M^*1$ or $M^*2$ described in the previous section).

Rule #5 uses $H_d$ as a figure of merit to rank order alternative batteries in terms of PCE. As McLaughlin et al. (1984) correctly realized, $H_d$ is not very useful by itself in the evaluation of alternative test composites that have a predetermined one-on-one match to jobs in the assignment process. An appropriate measure that estimates the OCE of sets of AAs, just as $H_d$ estimates the PCE of batteries (when LSEs are used as the composites), is much needed. We could suggest that our $M^*1$ or $M^*2$ may be used for this purpose, but they, like $M^2$, are too time consuming to compute as a substitute for a more valid approach. The best appraisal obtainable for the OCE of alternative composite sets appears to be the simulation methods of the type we will discuss in Chapter 4.

Rule #6 requires two steps: the subtraction of each row mean of $\mathbf{V}_a$ from the diagonal element in that row, and the computation of the variance within each row. It is desirable for the first value, the differences, to be positive and as large as possible. The row variance also should be as large as possible. Dependence on this rule, however, can result in the reduction of OCE in the AAs. This rule-of-thumb should not be used to estimate either OCE or PCE; it is likely that most users of this rule have it confused with rule #7, a much more useful rule.

Rule #7 substitutes columns for rows in R#6. The AA contributing the most to OCE may well be the one with the largest positive value when the column mean of $\mathbf{V}_a$ is subtracted from the diagonal element. The variance of the column adds little additional information regarding the OCE. However, column variance is most important in the evaluation of $\mathbf{V}$ in order to identify the test that may contribute the most PCE.

Rule #8 considers the column variance of $\mathbf{V}$ to be approximately proportional to PCE. It should be noted that R#7 and R#8 are fairly good approximations of the contribution that a single predictor makes to either OCE or PCE. When the contribution of two or more predictors to either OCE or PCE is to be estimated, the intercorrelations among predictors become important and thus an additional rule-of-thumb indicator should be used.

Rule #9 pertains to the separate dimensionalities of the sets of predictors and the criteria. The figure of merit for this rule-of-thumb evaluation is the number of orthogonal factors of a useful size that results from the factoring of either or both intercorrelation matrices. The practical figure of merit for R#9 is the dimensionality of $\mathbf{R}_t$ or $\mathbf{R}_a$ since knowledge of intercorrelations among job criteria is unlikely to exist, and the intercorrelations among components for the same job are not relevant. McLaughlin et al. (1984) reported that the number of common factors provided by an ASVAB correlation matrix was only four, of which only two had roots greater than one. These results were presented as a basis of their poor expectations for classification efficiency of aptitude areas drawn from this battery. The value of R#9 is that it establishes the upper bound of the figure of merit provided by the last rule, R#10.

Rule #10 relates to the number and magnitude of dimensions in the joint predictor-criterion space, the space spanned by the predicted performance measures. A factor solution of the matrix of covariances among these measures (one measure per job or job family) provides an estimate of the dimensionality of this space. The matrix to be factored has squared multiple correlation coefficients in the diagonals, as contrasted to the ones in the diagonals of $\mathbf{R}_t$ and $\mathbf{R}_a$ . The matrix representing the joint predictor-criterion space can be computed as either $\mathbf{V}(\mathbf{R}_t^{-1})\mathbf{V}'$ or as the reproduced matrix, $\mathbf{FF}'$, where $\mathbf{F}$ is the Dwyer factor extension solution (the extension of the complete factorization of $\mathbf{R}_t$ or $\mathbf{R}_a$ into the criterion space). The figure of merit for R#10 is defined rather vaguely as the number of dimensions in this space, usually as the number of orthogonal factors with roots over a specified size obtainable from a factorization of $\mathbf{FF}'$. Unique factors are accepted as adding to the dimensionality of the predictor-criterion space provided they have one validity coefficient of practical magnitude.

Brogden's (1959) model assumes that the dimensionality of the predictor-criterion space is one greater than the number of jobs. Since this assumption may never be met in practice for a set of more than a half dozen jobs, the robustness of the Brogden model with respect to this assumption should be determined by a simulation experiment.

The use of the above ten figures of merit as rule-of-thumb estimates of either PCE or OCE is particularly helpful in interpreting data provided by others. For example, Hunter (1986) uses R#6 to conclude that the OCE obtainable from the use of operational composites from the ASVAB is nil. We would not dispute a conclusion based on the use of R#7 that the indicated PCE is too low to justify the continued use of the existing AAs when used in unweighted standard score form to make or recommend assignments, even though we are unwilling to accept Hunter's conclusions based on the use of R#6 or his statement that the small deviation from undimensionality found in the $V_a$ type matrix provided by Hunter is the result of sampling error.

We believe research decisions should not rely on rule-of-thumb measures but should instead use the model sampling approach described in Chapter 4 or the simulation approach employed in Zeidner and Johnson (1989b). Results from use of either $M^*1$ or $M^*2$ for McLaughlin et al.'s (1984) data would provide very interesting estimates of the OCE obtainable from the alternative AA sets they evaluated using $M^2$. Further research, based upon either model sampling methodology or a simulation approach, using a large available data base, would be required to relate such a psychometric index to OCE in terms of MPP. The great value of simulation methods is that results can be directly expressed as an MPP standard score.

## F. THE INVESTIGATION OF CLASSIFICATION POLICY ISSUES BY THE SIMULATION OF ASSIGNMENT PROCEDURES

In 1968 an Army research team was assigned the responsibility of developing the capability of evaluating alternative personnel policies through the simulation of personnel operations. Two different approaches were incorporated in "Simulation Models for Personnel Operations" (SIMPO) (Olson, Sorenson, Haynam, Witt, and Abbe, 1969).[13] The better known approach used network flow models to track personnel through various types of assignments, training requirements, and promotions. We are more interested in the lesser known SIMPO entity models employed to evaluate selection and classification policies and procedures (Johnson and Sorenson, 1974).

SIMPO was an OR effort and most of its published reports were methodological in nature; the substantive results of the OR studies of personnel policies were usually not published. Fortunately a few of the methodology reports of SIMPO provided examples that bear on the relationship of data characteristics and classification processes to potential classification efficiency (PCE). A matured model sampling capability, much like the one described in Chapter 4, was described by Niehl and Sorenson (1968) as a "SIMPO I Entity Model for Determining the Quantitative Impact of Personnel Policies." A model was described that generates synthetic scores for hypothetical individuals (i.e., entities). The model used the entity scores as input into a simulated personnel system process reflecting

[13] SIMPO was a requirement in the Army Master Study program which was implemented as a BESRL (a predecessor of ARI) Work Unit, "Computerized Models for the Simulation of Policies and Operations of the Personnel Subsystem--SIMPO-1."

prescribed personnel policies and procedures. Most importantly, the model output at desired points in the simulated sequence of personnel actions was expressed as MPP standard scores. This model provided a valid and inexpensive capability for measuring the OCE or PCE resulting from alternative selection and/or classification policies.

Sorenson (1965a) simulated a mobilization population for a SIMPO model sampling experiment in which the gain in PCE provided by using LSEs instead of aptitude areas was evaluated. The means and covariances of the generated scores had expected values equal to those for the Army Classification Battery (ACB) tests in the mobilization population. Predicted performance scores were computed from full regression equations based on the population covariances. Separate validity vectors for eight job families were based on the validities of 55 jobs (MOS) corrected for restriction in range to provide estimates of job validities in the mobilization population. The effectiveness of eight two-test composites with weights of one or two were compared, as assignment variables used by an LP program, with the effectiveness of using full regression equations using all eleven tests in the ACB. The criterion variables for which validities were available were primarily Army school grades in an era when such grades were normative, reliable, truly indicative of the soldier's job knowledge, and became a permanent part of a soldier's record. The use of such school criterion variables typically provide more dimensionality in the predictor-criterion space and indicated greater PCE than does on-the-job criteria based on ratings. The validities of the two combat aptitude areas (AAs) were, however, computed only against criterion measures based on performance ratings of soldiers stationed in the continental United States.

Twenty entity samples of size 300, thirty samples of size 200, and two hundred samples of size 100 were generated. Appropriate quotas for each job family were used in conjunction with an LP program to assign the entities in each sample to one of eight job families. Assignment was accomplished once using the AAs as the assignment variables, and a second time using the full regression equations as the assignment variables. The MPP Army standard score was separately computed for each assignment procedure. The distributions of the MPP standard scores for the two assignment procedures did not overlap at all, even for the samples of size 100, and the Army standard score means (mean = 100 and SD = 20) were 103 when AAs were used and 107 when full regression equations were used as the assignment variables. The MPP Army standard score would have equalled 100 if random assignment had been used.

Thus, the gain over random assignment is roughly doubled by substituting full regression equations for the AAs. In contrast, McLaughlin et al. (1984) found, using $M^2$ as an estimate of differential prediction efficiency, that operational AAs were 64 percent as efficient as full regression equations computed separately for each of 98 jobs. The shredding out of job families into jobs and computing separate LSEs for each job was shown in unreported Army model sampling experiments to increase considerably the advantage of assignment by LSEs over the assignment by AAs. Thus the discrepancy between the results (i.e., the gain in MPP due to use of LSEs rather than aptitude areas) of Sorenson and McLaughlin, et al., would have been even greater if Sorenson had used separate LSEs for jobs instead of for job families. This discrepancy may be attributable to either differences in the PCE of the two test batteries, to the methodology for computing PCE, or, more likely, to both.

In 1965 the Army transformed each soldier's aptitude area score to single digit scores ranging from 0 to 9 in order to simplify operational assignment procedures. Model sampling experiments were conducted in which several alternative scales (including the operational 0-9 non-linear scale and an almost fully continuous range of scores) were evaluated to determine their contribution to PCE. Sorenson (1967) reported on a model sampling experiment in which entities were generated to have an expected predicted performance covariance matrix with equal off diagonal elements. Assignments were made using uniform quotas (0.0625 for 16 jobs) and again using perturbed quotas ranging from 0.0062 to 0.1187. Quotas were modified to provide for whole numbers of entities to be assigned to each job.

The results showed that scales with more intervals were generally superior to those with fewer intervals. Many different variables were considered. In one analysis, MPP results, after assignment, were compared for four combinations of two assignment variable scales and two selection policies. Selection policy A was to accept everyone with AFQT scores greater than 10 (expressed as a percentile score). Policy B was the same as policy A except that those with AFQT scores between 29 and 31 also had to have two LSEs above 90 (Army standard score). When the assignment variables are continuous full regression equations the MPP Army standard score is 106.49 under selection policy A and 107.80 under selection policy B. When the full regression equation scores are converted to a one-digit score and used as the assignment variables, the policy A MPP score is 105.83 and the policy B MPP score is 107.74. The scale effect is obviously trivial in the Policy B situation, but worth considering when selection is less restrictive.

Sorenson (1965) provides a different perspective of the importance of the gain provided by policy A over policy B (i.e., 106.49 vs. 107.80). He describes the following impact: "Under the conditions resulting in an allocation average of 106.49, a total of 660 men were assigned in jobs in which their expected performance was below 90 (expressed in Army standard scores); in comparison, 358 were thus assigned under conditions resulting in an allocation average of 107.80. This number represents a decrease of 46 percent assigned to a job for which performance is predicted to be low (90 or below)." (p. 3.) These results, along with the less dramatic increase of the number of high performers (those with predicted performance Army standard scores of 130 or above), led Sorenson to conclude: "These results may reasonably be generalized to the conclusion that an important increase in the number of outstanding performers and a reduction in the number of below-average performers may be achieved even though the increase in the allocation average is so small as to appear inconsequential." (p. 43.) Sorenson makes an important point, but we believe the comparison of policies A and B were not the best examples, since policy B requires that everyone have two scores above 90. No one would have had a score below 90 in the job to which an individual was assigned if the quotas had permitted each man to be assigned to one of his highest two scores.

Another member of the Army research team, Harris (1967), used the SIMPO model sampling design to evaluate the PCE of several pairs of batteries selected from a larger experimental test pool. In each pair of equal sized batteries, one was selected to maximize $H_a$ and the other to maximize $H_d$.

Twenty tests were sequentially selected by each method. School final course grades for 12 Army MOS were corrected for restriction in range to approximate validities for a mobilization population. Similarly the intercorrelations of 32 experimental tests were corrected to represent the same mobilization population. The intercorrelation matrix was computed on a sample of 2480 soldiers while the sample size for the Army school courses ranged from 103 to 305.

Assignments of synthetic individuals (entities) were accomplished using LSEs computed on the specified battery and using an LP program with uniform quotas for the 12 jobs. Assignments were evaluated using LSEs based on all 32 tests. These LSE scores for both assignment and evaluation were probably adjusted (the author does not indicate) to have means of 100 and standard deviations equal to the result of inserting tests (with means of 100 and standard deviations of 20) into raw score regression equations. In this way the MPP Army standard score after random assignment would equal 100.

The batteries of size 5, 10, and 20 selected to maximize $H_d$ were all superior to those selected to maximize $H_a$. All differences were statistically significant. For the batteries of size 5, with one overlapping test, the MPP scores were respectively 109.79 and 110.89, a 10 percent gain over random assignment provided through the use of $H_d$ instead of $H_a$.

Non-cognitive tests were selected early in the sequence using either $H_d$ or $H_a$. Arithmetic Reasoning, probably the purest measure of general mental ability for Army jobs, was the first to be selected to maximize $H_a$, and "verbal" (a vocabulary and reading comprehension test) was the last. Three of the best five tests selected to maximize $H_d$ were self description measures, the third to be selected was perceptual speed. Surprisingly, the fourth test selected in the sequential test selection against $H_d$ was verbal.

An existing sample of applicants or employees can be used as the source of predictor scores in lieu of the generation of entities by model sampling techniques. To conduct simulations for the evaluation of selection/classification policies, the computation of predicted performance measures for every job can be accomplished (as LSEs based on the total set of predictors), just as in the model sampling experiments described above. Alf and Wolfe (1968) conducted a similar simulation of Navy jobs during the same era as the Army was conducting the model sampling experimentation.

Five hundred eighty-seven complete data cases from 905 enlisted men who entered the San Diego Naval Training Center during a single week provided the predictor scores for this simulation. The assignment and evaluation variables compared in the simulation include the following: (1) AA scores; (2) school grades predicted from test scores; (3) a training course pass/fail criterion predicted from test scores; (4) training cost (without pay and allowances); (5) training cost (with pay and allowances); (6) manpower shortage in each rate (Navy equivalent of an MOS); and (7) criticality (the product of a school criticality index and the value of the third measure listed above).

Assignments were made using an LP program to optimize, in turn, each of the above 7 variables. For each of these seven assignments to jobs, plus one accomplished using the operational (hand) method, and another by random assignment, the results were evaluated using every assignment variable as an evaluation variable. As one would expect, an optimal assignment algorithm yielded the best mean performance score when the evaluation variable and the assignment variable were the same, but the third evaluation variable described above when used as the assignment variable, was also second best when evaluated on each of the other variables (with the exception of the AAs); the AAs are however, undoubtedly of minimal appropriateness as an evaluation measure.

The regression equation yielding predicted pass/fail in school courses also predicts the school grades almost as well as does the regression equation developed to predict grades. In addition, the predictor of pass/fail is uniformly better for lowering costs and increasing the manning level index as compared to the predictor of grades. Thus the authors recommend the predictor of pass/fail as the assignment variable with the highest across-the-board utility.

Three of the other evaluation measures, (4), (5) and (7) listed above, include predicted pass/fail (Ps) as an ingredient. One wonders, if predicted grades had been substituted for Ps in those formulae, whether predicted grades would have replaced Ps as the second best assignment variable for these three evaluation criteria. Also, the metrics on which the assignment results are expressed are not comparable. We cannot judge whether the larger gain in percentage predicted to succeed provided by the pass/fail predictor is really of greater utility than the gain in predicted school grades provided by the predictor of grades.

Despite our skepticism as to the meaningfulness of their recommendations, we believe this study was outstanding for its era. The seventh evaluation variable, which combined a job criticality index with a measure of predicted P/F, deserves further consideration, and possible emulation (of course, substituting predicted performances for predicted P/F) by research personnel.

## G. ESTIMATING PCE BY APPLYING META ANALYSIS DATA TO MODELS OF THE NATIONAL ECONOMY AND THE NAVY

A different approach to the estimation of the selection/classification efficiency obtainable from a battery is provided by Hunter and Schmidt (1982) and Schmidt, Hunter and Dunn (1987). The first defines a hypothetical three-test battery and a model of the national economy. The second applies a two-test battery to a model of the Navy. Both models abstract all features of the target organization essential to the computation of PUE under their assumptions. The major, more general, assumption common to both studies relates to the preeminent role ascribed to general mental ability in the prediction of job performance and the use of job complexity categories. The authors assume that there is one dominant general factor in the space spanned by predicted job performance: a unidimensional performance measure (not to be confused with a general mental ability factor in the predictor space) that accounts for all demands made on cognitive ability. A corollary to this assumption is that the potential for predicting this primary performance factor lies in a single measure--general mental ability--and that there are only one or two reliably identified additional dimensions in the joint predictor-criterion space--psychomotor ability and possibly perceptual speed. Also linking the two studies is the use of the Dictionary of Occupational Titles (DOT) information on jobs and the results of the U. S. Employment Service General Aptitude Test Battery (GATB) validity research.

These two studies differ from the studies described in the previous section in that GATB results rather than military results provide the estimates of intercorrelations and validities With respect to the Schmidt et al. (1987) model, generalized data from the GATB were used in the Navy study, rather than actual military empirical data in determining intercorrelations and validities.

The assignment techniques used in the studies in both sections are readily relatable. For example, assuming their numerical computations were accomplished correctly, the results of the Hunter and Schmidt (1982) study are the same as if the authors had generated synthetic scores to yield an expected correlation matrix equal to the one they stipulate, and then assigned these entities to jobs using a primal LP program. Similarly, assuming no computational errors, the Schmidt et al. (1987) study would have obtained the same results if they had used any one of the many off-the-shelf primal LP programs to make assignments, and used readily obtainable column constants to reject a percentage of the applicant population. Instead, they used a dual LP program (a modification, of which there are many, of the Brogden-Dwyer optimal regions algorithm) to effect optimal assignments and rejections. Our attention is centered on assumptions regarding the models of the respective organizations and the characteristics of the predictor batteries, rather than on how the effects of selection and assignment were determined.

Both the 1982 and 1987 studies rely on models of organizations in which assignments are being made. The models, either of the national economy (1982) or of the Navy (1987), consist of a description of job categories, and, separately by job category, the number of individuals and the value of each individual's productivity. Major categories were formed to maximize the credibility of the assumption that the jobs in each category are homogeneous with respect to both their validities and the value of their output, while keeping the categories few in number. In the Navy model, however, the validities of predictors varied across job categories as a function of a unidimensional job characteristic, job complexity.

As described in Zeidner and Johnson (1989a), Hunter and Schmidt (1982) assumed equal validity of general ability for all categories of jobs, specifically, product moment correlation coefficients of 0.40, but assumed differential validities for both spatial ability and perceptual speed. (i.e., validities are either 0.40 or 0.00, a detail which is not obvious from Figure 4.1). Perceptual speed ability was assumed to have a validity (product moment correlation coefficient) of 0.40 with the "performance in skilled trades" category, and zero validity for performance in all other categories. Similarly, psychomotor ability was assigned a validity of 0.40 for the "clerical" category and zero validity for performance in all other categories. General ability was assumed to correlate 0.40 with each of the other two abilities; it was assumed that these other two abilities provided half of their predictive capability because of their general mental ability content, and were given a correlation coefficient between the two of 0.16 to reflect the assumption that their non-zero correlation was entirely due to their general mental ability content.

The effects of three alternative assignment processes on utility were computed for the model of the national economy. These three assignment modes were: (1) a random process, (2) a hierarchical classification process using only general ability as a univariate assignment variable, and (3) an optimal assignment process with separate two variable LSEs used for all but one of the aggregated job categories.

The author's assignment process for univariate hierarchical classification was essentially the same as the one used in our example in Chapter 1. However, in our example, the differences in MPP scores across jobs were due entirely to differences in validity, while in the national economy model these differences are due entirely to the differential values placed on the productivity of job categories.

To accomplish the desired optimal assignments using only one assignment variable--general mental ability--the population need only be ranked on general mental ability and those with the highest general mental abilities assigned to the most highly valued job, the next highest block on general mental ability assigned to the next highest valued job, etc., until all jobs are filled. These same assignments would be made if an LP program were used instead of this very simple process, when there is only one assignment variable and differential values are specified for jobs.

It is frequently enlightening to visualize a computational process in terms of a simulation. If scores necessary to conduct a simulation were obtained, optimal assignment of the population to the job categories could be accomplished by computing predicted performance in terms of the LSEs for each job category, weighting LSEs by job value, and using a LP program to assign individuals to meet the quotas for each category; the objective would be to maximize the value weighed predicted performance for each job category. It should be noted that even for the job categories that have only non-zero validities for general mental ability, the LSEs would have non-zero negative weights for both the psychomotor ability measure and the perceptual speed measure (two very efficient suppressor variables are present).

This approach could have been used for a sample of synthetic entities generated in the manner discussed in Zeidner and Johnson (1989a). Hunter and Schmidt used a numerical solution of this same problem expressed as definite integrals of normal curve functions to arrive at a solution that should provide the same utility score as would result from the alternative simulation process defined above. Either solution should provide the same value for classification efficiency; for these two studies PCE would be expressed in terms of utility instead of a MPP standard score.

The Navy model used in Schmidt et al. (1987) is very similar to the national economy model study in that both studies model a system in terms of job categories. For each of these job categories the authors stipulate the number of incumbents, the value of the production of an individual, and the predictability of an individual's performance. The Navy model differs from its predecessor in the following ways: (1) jobs are categorized on a continuum of complexity, instead of by traditional major job families, (2) more realistic estimates of test validities for each job category are provided, (3) the perceptual speed ability is either omitted or combined with general ability, and (4) a rejection category is included. We will discuss each of these differences.

Hunter (1980) in a study based on the meta-analysis of GATB data concluded that this battery tapped three abilities, essentially the three utilized in the national economy model. After classifying the jobs of the DOT titles into one of five complexity levels, he concluded that general mental ability and psychomotor ability were complementary, providing essentially equal validity for the combined measures across all but clerical jobs.

Complexity, defined as the level of cognitive information processing demands of a job, is claimed to require more general mental ability and less psychomotor ability at the high end of the complexity scale, and vice versa at the low end of this scale. The identification of a Navy job's location on this continuum, by first matching the Navy job with a DOT job that has a tabled complexity value, permits a conversion that provides the validities for both general mental and psychomotor ability. Jobs not convertible to a DOT complexity level were assigned to a complexity level by judgment, and, through their membership in the job group having a given complexity, received an estimate of zero order validities of both ability measures against performance on the job. Since the correlation between those two ability measures was assumed to be 0.35, the multiple correlation coefficients are readily computed. These latter coefficients, using validities first estimated for the GATB and then adjusted for the ASVAB, yield the following multiple correlation coefficients (the highest of the five complexity levels listed first): 0.64, 0.65, 0.59, 0.54, 0.49.

While there are interesting theoretical implications associated with the complexity continuum, the advisability of modeling the Navy in terms of complexity levels, instead of more traditional job families, hinges on three practical considerations. The first is whether validities are more homogeneous for jobs clustered on this continuum than would be provided by alternative clustering techniques. The second is whether the validities of jobs are better estimated by their identification with a complexity level than by alternative categorizations. The last is whether jobs can be as objectively classified into complexity levels as into the more traditional groupings.

Results for selection and classification using the Navy model were provided separately for the univariate hierarchical classification mode and the two variable classification modes that depend primarily on hierarchical layering as the source of most of the added classification efficiency. The univariate selection-assignment mode results were reported separately for: (1) a purer form of general mental ability, and (2) a general mental ability measure augmented by a clerical speed measure to form a single operational test.

The authors apparently believe that the perceptual speed ability, if measured by a separate test and included in the battery, would have added some PCE to such a 3-test battery, as compared to the PCE in the two-test battery used in conjunction with the Navy model. Their rationale for this treatment of clerical speed was that this ability is currently in the ASVAB and thus cannot be presented as a potential augmentation of the battery, and that much of the contribution of a clerical speed test to selection/classification could be captured in the Navy model through the combining of general mental ability and clerical speed measures. It appears likely that the deleterious effect that the addition of a third assignment variable would have had on the usefulness of their analysis procedure was also an important motivator in the making of this decision.

The results reported for the successive upgrading of the test battery is in terms of dollar value productivity that potentially can be provided by optimal selection and classification. The assignment variables to be maximized in the assignment process are value-weighted predicted performance scores and the assignment process is equivalent to an LP program. It is not feasible to convert these results to MPP scores comparable with the results of other studies.

The interpretation of the results expressed in utility terms requires the consideration of the assumptions and procedures involved in arriving at the dollar value and the spread of productivity for jobs at different levels of complexity. Other controversial issues that could affect the dollar value of results include the use of an "equilibrium model" instead of a "cohort model" and their methods for handling of costs. A discussion of these issues is beyond the scope of this chapter.

When utility is expressed in terms of gain over random selection and assignment, the gain provided by use of general mental ability alone is reported to be 15.07 percent. Changing the basis of comparison, the gain over general mental ability (as the surrogate for the ASVAB) by augmenting the general mental ability measure with perceptual speed to form a single test is 3.19 percent, and the gain over general mental ability provided by a two-variable optional assignment process, using the GATB psychomotor test as the second variable, is 5.20 percent.

The method used to reject applicants appears to have an equivalent objective to our MDS process. According to the authors, "The optimal assignment is to reject those whose productivity would have been least. This can be done by 'adjusting' performance scores in the reject condition so that selection of those with highest adjusted performance scores will place the correct applicants in the reject group....these will be the workers for whom there is least loss if they are assigned to the reject category. If the adjustment coefficient for each reject category is set correctly, the necessary number of recruits will be assigned to that category in an optimal manner." (p. 66.)

While it appears to us that the authors have the correct idea of how to reject those who would perform the worst if accepted and optimally assigned, they should not have searched for adjustment coefficients for the reject "jobs." The adjustment coefficients (what we call the job or column constants in Chapter 1) corresponding to each job are the same for both selection and classification; the additive job constant used to optimize classification is the same as the one which will optimize multidimensional selection. Once the appropriate job constant is added to each predicted performance, each applicant should be tentatively assigned to his highest adjusted score, and enough of those with the higher adjusted scores selected to fill the quotas; those remaining after the quotas are filled are those that should be rejected.

It would be interesting to consider the usefulness of applying the Schmidt et al. (1987) Navy model approach to the Army. The Army could undoubtedly also be modeled in terms of the complexity continuum. Using the validity data provided by McLaughlin et al. (1984, p. 22), and applying rule-of-thumb principle R#7, we see that good differential validity is, when considered separately, provided by two of the nine aptitude areas (AAs), the other six AAs have validities against the corresponding job families that are lower than their mean validities for the non-corresponding jobs.

The two AAs showing good differential validity (*CL* and *ST*) had their validities computed on samples of 10,368 and 7,061, respectively; the smallest of the other job families still had an *N* of 2,571. This configuration of results is confirmed in other samples and we believe it would be hard to argue convincingly that the PCE indicated by this data is based on error and that there are only two abilities measured by the ASVAB, general mental ability and clerical speed; there is at least one ability tapped in the joint predictor-criterion space for Army jobs in addition to general mental ability and clerical speed.

Accordingly, we believe the Army model would have to represent the ASVAB by no fewer than three ability measures, each measure consisting of a composite that may contain several tests; for the purpose of the model it makes no difference whether a composite consists of one test or many. The PCE and corresponding utility for that three-test battery would provide the base line against which the PCE and eventually utility for the same battery augmented by a fourth test, the GATB psychomotor test, could be compared. We would be skeptical of the meaningfulness of finding out what a psychomotor test would add to only a single general mental ability test.

In summary, we find the two studies described in this section to be important additions to the literature on the contribution of classification to utility. However, we would not recommend that decisions concerning the value of the GATB psychomotor test be based on these two studies. Questions about the basic assumptions need more complete answers and the representation of the ASVAB by more than one measure needs to be incorporated in the Navy model before the value of additional tests for Navy classification can be realistically estimated. Furthermore, all the utility results of the study are highly dependent on the use of value-weighted assignment variables in the selection and classification process. All of the gain over random classification from using a single predictor (general mental ability) in the Navy model would vanish if test composites used in the assignment process were not value weighted. Since such value weighting implies the making of major policy decisions on quality distribution involving considerable organizational sensitivity, the results from the Navy model with all such value weights equal to one might be of more relevance to decisionmakers.

# APPENDIX 2A
# BASIC CONCEPTS AND NOTATION

## APPENDIX 2A.1: INTRODUCTION

The technical appendices for this chapter and the following chapters use a consistent matrix notation. All later appendices, except where specifically noted, will build upon concept development and derivations presented in earlier appendices. The order of presentation is a compromise between the occurrence of concepts in the text and the need for a sequential presentation of technical concepts. Once notation and/or concepts have been presented we will freely use them thereafter.

## APPENDIX 2A.2: SOME FREQUENTLY USED MATRICES

All matrices are designated by capital letters. Capital letters always indicate a matrix except that R is occasionally used to represent a multiple correlation coefficient and S has been used to represent a standard deviation. A capital letter without a subscript represents a class of matrices; a subscripted matrix stands for a specific type of matrix within its class. An explanation in the text may sometimes take the place of a subscript.

A standard notation for dimensions is used to describe matrices. The first dimension describes the number of rows and the second dimension the number of columns. Commonly used matrices are as follows:

**Y** = an *N* by n matrix of standardized predictor (test) scores; underlining indicates that each score in the matrix is divided by the square root of *N*; for example, $\underline{\mathbf{Y}}'\underline{\mathbf{Y}} = \mathbf{R}_t$.

$\mathbf{Z}_u$ = an *N* by *m* matrix of standardized criterion scores; underlining indicates that each score is divided by the square root of *N*; for example: $\underline{\mathbf{Y}}'\underline{\mathbf{Z}}_u = \mathbf{V}'$, $\underline{\mathbf{Z}}_u'\mathbf{Y} = \mathbf{V}$, $\underline{\mathbf{Z}}_u'\underline{\mathbf{Z}}_u$ is usually unknown but may be hypothesized in some model sampling experiments.

**Z** = an *N* by n matrix of predicted performance (PP) scores the standard deviation of these scores is equal to the correlation of the PP variables with the corresponding criterion variables; $S_p^{-1/2}\,\underline{\mathbf{Z}}'\underline{\mathbf{Y}} = \mathbf{V}$, $\underline{\mathbf{Z}}'\underline{\mathbf{Z}} = \mathbf{C}_p$.

**Q** = an *N* by *k* matrix of factor scores.

**R** = matrices of correlation coefficients with ones in the diagonals.

$\mathbf{R_t}$ = $n$ by $n$ matrix of correlation coefficients among predictors (usually selection or classification tests).

**V** = $m$ by $n$ matrix of validity coefficients (correlations between n predictor variables and $m$ job criterion variables); $\mathbf{Z}_u'\mathbf{Y} = S_p^{-1/2}\mathbf{Z}'\mathbf{Y} = \mathbf{V}$.

**C** = covariance matrix with variances in the diagonals.

**S** = diagonal matrix of variances (e.g., the diagonal elements of **C**). $\mathbf{C_p} = \mathbf{S_p}^{1/2}(\mathbf{R_p})\,\mathbf{S_p}^{1/2}$.

$\mathbf{C_p}$ = $m$ by $m$ covariance matrix; the covariances among predicted performance estimates; the diagonal elements are multiple correlation coefficients; $\mathbf{C_p} = \mathbf{V}\,(\mathbf{R_t})^{-1}\,\mathbf{V}'$.

**F** = factor solutions in matrix form; the elements are regression weights applied to the column variables in standard score form to estimate the dependent variables represented by the rows. **FF'** equals or approximates either an **R** or **C** matrix.

**A** = eigen vector matrices; $\mathbf{A'A} = \mathbf{I}$ and, if a square matrix $\mathbf{AA'} = \mathbf{I}$; if **A** is a rectangular, orthonormal matrix, $\mathbf{AA'} = \mathbf{I}$ and $\mathbf{A'A}$ does not equal **I** but is idempotent.

**D** = eigen value matrices; diagonal matrices such that $\mathbf{ARA'} = \mathbf{D}$, $\mathbf{ACA'} = \mathbf{D}$, etc.

**T** = transformation matrices such that $\mathbf{R_tT} = \mathbf{F_t}$, $\mathbf{VT} = \mathbf{F_t}$, or $\mathbf{FT_r} = \mathbf{F_r}$.

$\mathbf{F_t}$ = an orthogonal factor solution of $\mathbf{R_t}$, thus $\mathbf{F_tF_t}' = \mathbf{R_t}$ (an $n$ by $n$ or an $n$ by $k$ matrix, $k < n$).

$\mathbf{F_v}$ = an orthogonal factor solution such that $\mathbf{F_vF_v}'$ approximates or equals $\mathbf{C_p}$; $\mathbf{F_v}$ is a factor extension of $\mathbf{F_t}$ into the joint predictor-criterion space (an $m$ by $n$ or an $m$ by $k$ matrix, $k < n$, with the number of factors, $k$ or $n$, equal to the corresponding $\mathbf{F_t}$).

$\mathbf{F_c}$ = a principal component factor solution of $\mathbf{C_p}$; $\mathbf{F_cF_c}' = \mathbf{C_p}$ and $\mathbf{F_c}'\mathbf{F_c} = \mathbf{D_c}$, where $\mathbf{D_c}$ is a diagonal matrix of eigen values; $\mathbf{F_c} = \mathbf{A_cD_c}^{1/2}$.

$\mathbf{F_p}$ = an orthogonal factor solution derived as an orthogonal transformation of $\mathbf{F_v}$, which equals $\mathbf{F_c}$, (assuming that null factors are equivalent to "no" factors);when $\mathbf{F_t}$ has $n$ columns, and $m > n$, $m - n$ columns of $\mathbf{F_c}$ will be all zeros while $\mathbf{F_p}$ will have only $n$ columns; when $n > m$, $\mathbf{F_c}$ will have only $m$ columns while $\mathbf{F_p}$ will have $(n - m)$ null, all zero, columns.

**H** = an $m$ by $m$ matrix in which each element is equal to $(1/m)$.

$\mathbf{G} = (\mathbf{F_V} - \mathbf{HF_V})$.

$H_a = \mathrm{tr}\,(\mathbf{F_V F_V}') = \mathrm{tr}\,(\mathbf{F_V}'\mathbf{F_V})$; Horst's "absolute validity" index, a measure of selection efficiency for FLS composites in a multi-job/criterion situation.

$H_d = \mathrm{tr}\,(\mathbf{G}'\mathbf{G}) = \mathrm{tr}\,(\mathbf{GG}')$; Horst's "differential validity" index, an estimate of potential classification efficiency.

Note that all factor solutions used in these appendices are in either total test space or joint predictor-criterion space; no solutions in common factor space will be utilized. We will not make the distinction between factor analysis and component analysis sometimes made by investigators in order to emphasize the differences between the use of common factor space and total test space.

## APPENDIX 2A.3: SUPERMATRICES

The matrix $\mathbf{R_t}$ bordered below by the matrix $\mathbf{V}$ forms a $m + n$ by $n$ supermatrix denoted as

$$\begin{bmatrix} \mathbf{R_t} \\ \cdots \\ \mathbf{V} \end{bmatrix},$$

and

$$\begin{bmatrix} \mathbf{R_t} \\ \cdots \\ \mathbf{V} \end{bmatrix} \mathbf{T} = \begin{bmatrix} \mathbf{F_t} \\ \cdots \\ \mathbf{F_v} \end{bmatrix} = \mathbf{F} \quad .$$

Also,

$$\mathbf{FF}' = \left[\begin{array}{c:c} \mathbf{R_t} & \mathbf{V}' \\ \hdashline \mathbf{V} & \mathbf{C_p} \end{array}\right] \quad .$$

Note that $\begin{bmatrix} \mathbf{R}_t \\ \cdots \\ \mathbf{V} \end{bmatrix}$ can be thought of as an oblique factor solution in which the column variables (all tests) are oblique factors, and the row variables are the independent variables (the variables that load on the factors).

Using $\mathbf{T}$ as the means of transforming this oblique factor solution into an orthogonal solution, as indicated above, it is useful to define $\mathbf{T}^{-1}$ as a matrix whose elements are the cosines of the angles between the row variables (the orthogonal factors) and the column variables (the oblique factors). A square matrix, $\mathbf{F}_t$, comprising a complete factor solution of $\mathbf{R}_t$ will have the sums of squares for each row equal to one and column variables, orthogonal factors, with designated standard deviations of one. The elements of such a solution can be considered to be the cosines of the angles between vectors (in n space) representing the row variables (the predictors or oblique factors) and the orthogonal factors. Thus $\mathbf{F}_t\mathbf{T}^{-1} = \mathbf{R}_t$ and it follows that $(\mathbf{R}_t)\mathbf{T} = \mathbf{F}_t$. When the particular factor solution $\mathbf{A}_t\mathbf{D}_t^{1/2}$ is chosen to be equal to $\mathbf{F}_t$, we find that $\mathbf{T} = \mathbf{A}_t\mathbf{D}_t^{-1/2}$.

The same logic that calls for $(\mathbf{R}_t)\mathbf{T}$ to equal $\mathbf{F}_t$ is equally applicable to the relationship displayed by $\mathbf{VT} = \mathbf{F}_v$, and the columns of $\mathbf{F}_v$ represent the same variables, i.e., the same factors, as the columns of $\mathbf{F}_t$. Those factors are defined in terms of predictor (test) variables only; the criterion variables that represent the rows of $\mathbf{F}_v$ are correlated with factors that are defined entirely in terms of the test variables. Thus, $\mathbf{F}_v$ can be thought of as the extension of the factors defined in test space into the criterion space. Defined in test space and extended into criterion space they can be said to span a joint predictor-criterion space.

## APPENDIX 2A.4: IMPORTANT RELATIONSHIPS AMONG THE DEFINED VARIABLES

A number of relationships among the matrices defined above are important to later developments. A number of these that occur most frequently are given below:

$\mathbf{R}_t = \mathbf{A}_t\,\mathbf{D}_t\,\mathbf{A}_t'$.

$\mathbf{F}_t = (\mathbf{R}_t)^{1/2} = \mathbf{A}_t\,(\mathbf{D}_t)^{1/2}\,\mathbf{A}_t'$, a Grammian factoring of $\mathbf{R}_t$.

$\mathbf{F}_t = \mathbf{A}_t\,\mathbf{D}_t^{1/2}$, when defined as a principal component (PC) solution of .

$\mathbf{Q_t}$ $= \mathbf{Y}\ \mathbf{R}^{-1}\ \mathbf{F}_t$. an $N$ by $n$ matrix of factor scores corresponding to a specific $F_t$.

$\mathbf{F}_t$ $= \underline{\mathbf{Y}}'\underline{\mathbf{Q}}_t$.

$\mathbf{F}_v$ $= \mathbf{S}_p{}^{-1/2}\underline{\mathbf{Z}}'\underline{\mathbf{Q}}_t$.

$\mathbf{T}$ $= \mathbf{A}\ \mathbf{D}^{-1/2}\ \mathbf{A}'$, if $\mathbf{F}_t = \mathbf{A}\ \mathbf{D}^{1/2}\ \mathbf{A}'$, a Grammian factor solution.

$\mathbf{T}$ $= \mathbf{A}\ \mathbf{D}^{-1/2}$, if $\mathbf{F}_t = \mathbf{A}\ \mathbf{D}^{1/2}$, a principal component (PC) solution.

$\mathbf{T}$ $= (\mathbf{F}_t'\mathbf{F}_t)^{-1/2}\ \mathbf{F}_t'$. for any $\mathbf{F}_t$ such that $\mathbf{F}_t\mathbf{F}_t' = \mathbf{R}_t$; (note that this is the Dwyer (1937) formula for $\mathbf{T}$).

## APPENDIX 2.B
## THE JOINT PREDICTOR-CRITERION SPACE AND THE FACTOR EXTENSION PROCESS

We have denoted an $N$ by $n$ matrix of factor scores divided by the square root of $N$ as $\underline{\mathbf{Q}}_t$ and noted that $\mathbf{F}_t = \underline{\mathbf{Y}}'\underline{\mathbf{Q}}_t$ and $\mathbf{F}_v = \mathrm{S}_p^{-1/2}\underline{\mathbf{Z}}'\underline{\mathbf{Q}}_t$. The concept of factor extension requires the definition of $\mathbf{Q}_t$ in terms of the predictor variables, that is $\mathbf{Y}$, and the obtaining of the correlations of the criterion variables with these same factor scores. Describing $\mathbf{Q}_t$ in terms of $\mathbf{Y}$ we have $\mathbf{Q}_t = \mathbf{Y}\ \mathrm{R}_t^{-1}\mathrm{F}_t$, and

$$\mathbf{F}_v = \mathrm{S}_p^{-1/2}\underline{\mathbf{Z}}'\underline{\mathbf{Q}}_t = \mathrm{S}_p^{-1/2}\underline{\mathbf{Z}}'\underline{\mathbf{Y}}\ \mathrm{R}_t^{-1}\mathbf{F}_t = \mathbf{V}\mathbf{R}_t^{-1}\mathbf{F}_t.$$

Assuming $\mathbf{F}_t = \mathbf{A}_t\mathbf{D}_t^{1/2}$ and noting that

$$\mathbf{R}_t^{-1}\mathbf{F}_t = \mathbf{A}_t\mathbf{D}_t^{-1}\mathbf{A}_t'\mathbf{A}_t\mathbf{D}^{1/2} = \mathbf{A}_t\mathbf{D}_t^{-1/2},\ \mathbf{F}_v = \mathbf{V}\mathbf{A}_t\mathbf{D}_t^{-1/2},$$

we see the same expression we obtained using either our "$\mathbf{T}$" approach or Dwyer's formula.

An investigator may choose to first compute $\mathbf{F}_t$, rotate to simple structure and then extend the rotated solution to the criterion variables. Alternatively he/she may wish to factor $\mathbf{C}_p$, i.e., compute $\mathbf{F}_c$, rotate to a meaningful solution, and then extend to the predictors, permitting definition of the rotated factors in terms of the better understood selection-classification test variables.

If we wish to start with a PC solution of $\mathbf{R}_t$, $\mathbf{T}$ is equal to $\mathbf{A}_t\mathbf{D}^{-1/2}$ and we see that $\mathbf{F}_t = \mathbf{R}_t\,\mathbf{T} = \mathbf{A}_t\,\mathbf{D}_t^{1/2}$, and $\mathbf{F}_v = \mathbf{V}\mathbf{T} = \mathbf{V}\mathbf{A}\mathbf{D}_t^{-1/2}$. After the rotated solution is obtained as $\mathbf{R}_t\mathbf{A}\mathbf{D}_t^{1/2}\mathbf{T}_r$, the rotated solution in the joint predictor-criterion space is $\mathbf{V}\mathbf{A}_t\mathbf{D}^{-1/2}\,\mathbf{T}_r$. Note that if we substitute $\mathbf{F}_t = \mathbf{A}_t\mathbf{D}_t^{1/2}$ in the Dwyer formula for $\mathbf{T}$ we see that it simplifies to $\mathbf{T} = \mathbf{A}_t\mathbf{D}_t^{-1/2}$, just as we would expect.

Commencing with $\mathbf{F}_c$, a PC solution of $\mathbf{C}_p$, and rotating to simple structure in terms of the criterion variables, resulting in $\mathbf{F}_c\,\mathbf{T}_r$, the investigator would wish to extend his rotated solution to the predictor space. This extension process would commence by finding $\mathbf{T}_p$ such that $\mathbf{V}\mathbf{T}_p = \mathbf{F}_p = \mathbf{F}_c$. The rotated solution for the predictor variables would then be $\mathbf{R}_t\mathbf{T}_p\mathbf{T}_r$. The required $\mathbf{T}_p$ can be written as $\mathbf{A}_t\mathbf{D}_t^{-1/2}$ postmultiplied by the eigen vectors of $(\mathbf{F}_v'\mathbf{F}_v)$.

$\mathbf{F}_v$ computed as the factor extension of $\mathbf{F}_t$ can be transformed into a PC solution of $\mathbf{C}_p$ by finding an orthogonal transformation matrix $\mathbf{A}_p$, such that $\mathbf{F}_v\ \mathbf{A}_p = \mathbf{F}_p$. This $\mathbf{F}_p$ must have the characteristics of a PC solution. That is, $\mathbf{A}_p'\mathbf{F}_v'\mathbf{F}_v\mathbf{A}_v = \mathbf{D}_p$, where $\mathbf{A}_p'\mathbf{A}_p = \mathbf{I}$, $\mathbf{A}_p\mathbf{A}_p'$ is an idempotent matrix and $\mathbf{D}_p$ is a diagonal matrix. It is well known that the matrices that will result in the diagonalization of a Grammian matrix, $\mathbf{M}$, in accordance with the equation $\mathbf{A}'\mathbf{M}\ \mathbf{A} = \mathbf{D}$ is unique and must be the eigen vectors and eigen values of $\mathbf{M}$. Thus, there can be only one orthogonal transformation of $\mathbf{F}_v$ that exhibits this rotation of $\mathbf{F}_p'\mathbf{F}_p$ by an orthonormal matrix and its transpose into a diagonal matrix--$\mathbf{F}_v\ \mathbf{A}_p$ must be the PC solution, $\mathbf{F}_p$. The desired expression for $\mathbf{F}_{tr}$ is seen to be as follows: $\mathbf{F}_{tr} = \mathbf{R}_t\ \mathbf{A}_t\ \mathbf{D}_t^{-1/2}\ \mathbf{A}_p\ \mathbf{T}_r$ which corresponds to the solution in the joint space of $\mathbf{F}_{vr} = \mathbf{V}\mathbf{A}_t\mathbf{D}_t^{-1/2}\mathbf{A}_p\mathbf{T}_r$.

# APPENDIX 2C
# THE USE OF TRIANGULAR FACTORS IN TEST SELECTIONS

The accretion method of sequential test selection to maximize the prediction of a single criterion commonly uses a triangular factor solution of the candidate tests extended to a single criterion variable. Horst (1955) adapted this well known test selection approach to multiple criteria, that is, to maximize the sum of the squared multiple correlation coefficients of the selected tests against performance in more than one job. Although the publishing date for the selection method that maximizes $H_a$ is a year later than for the method that maximizes $H_d$, it is clear that the use of $H_a$ constitutes a relatively minor generalization of the traditional accretion test selection method--as compared to the replacement of $H_a$ by $H_d$ as the figure of merit in the selection of tests.

In describing his sequential test selection method for maximizing $H_d$, Horst (1954) makes reference to Dwyer (1951) as the source of a method that is essentially a square root or triangular factorization of $\mathbf{R}_t$ extended to $\mathbf{V}$. Dwyer's computing algorithm was designed for implementation on the desk calculator and in this computer dominated age is of little interest. However, the concept of the Gauss-Doolittle triangular factorization method remains an important one. We provide a discussion of an example in this appendix and present a detailed algorithm for test selection that utilizes triangular factorization in Appendix 3C.

The first three tests selected in accordance with a prescribed figure of merit are depicted in the following triangular factor solution in which the factor loadings are written as semi-partial correlation coefficients. These three factors are readily extended to the remaining test variables and to the criterion variables using the same computational approach. The following example shows three triangular factors extended to three additional test variables and to three criterion variables.

We can depict a three factor triangular solution in which the first factor, $L_1$, corresponds to the first selected factor, the second factor ($L_{2.1}$) is the component of the second selected variable that is orthogonal to the first variable after adjustment to unit length. Similarly, the third factor ($L_{3.12}$) is the component of the third selected predictor variable orthogonal to both the first and second variables; each variable or variable component representing a factor is adjusted to unit length (i.e., has a standard deviation of one).

We depict the loading of the $i^{th}$ variable on Factor $L_1$ as $r_{i1}$, on $L_{2.1}$ as $r_{i(2.1)}$, and on factor $L_{3.12}$ as $r_{i(3.12)}$. In the following example, the first three rows (i.e.,$\mathbf{F}_{t1}$), represent the predictor variables selected to define the factors; rows 4 through 6 represent the remaining predictor variables, and rows 5 through 9 represent the criterion variables.

<u>EXAMPLE</u>

$$
\begin{array}{c}
\begin{array}{ccc} L_1 & L_{2.1} & L_{3.12} \end{array} \\
\begin{pmatrix}
1.0 & 0.0 & 0.0 \\
r_{21} & 1.0 & 0.0 \\
r_{31} & r_{3(2.1)} & 1.0 \\
\ldots & \ldots & \ldots \\
r_{41} & r_{4(2.1)} & r_{4(3.12)} \\
r_{51} & r_{5(2.1)} & r_{5(3.12)} \\
r_{61} & r_{6(2.1)} & r_{6(3.12)} \\
\ldots & \ldots & \ldots \\
r_{e1} & r_{e(2.1)} & r_{e(3.12)} \\
r_{f1} & r_{f(2.1)} & r_{f(3.12)} \\
r_{g1} & r_{g(2.1)} & r_{g(3.12)}
\end{pmatrix}
\end{array}
=
\begin{pmatrix}
\\
\mathbf{F}_{t1} \\
\\
\ldots \\
\\
\mathbf{F}_{t2} \\
\\
\ldots \\
\\
\mathbf{F}_{v} \\
\\
\end{pmatrix}
$$

If each variable takes its turn as the last column of the triangular matrix $\mathbf{F}_{t1}$, this augmented factor matrix then extended to the criterion variables to become $\mathbf{F}_v$, and the variance of the last column of $\mathbf{F}_v$ computed, the predictor variable contributing the least to $H_d$ can be identified. The coefficients in this last column of $\mathbf{F}_v$ are the regression weights appropriate for application to the predictor component that is orthogonal to all of the other predictors remaining in the pool of predictors. These coefficients are equivalent to the elements of **W**, i.e., regression weights, that Brogden (1959) notes make no contribution to classification efficiency when the weights are essentially equal across jobs. Brogden's method for selecting tests for elimination is discussed further in Appendix 3A.

The algebraic equivalent to computing $m$ separate triangular $\mathbf{F_v}$ solutions, each solution placing a different variable in last place, is more economically obtained by using Horst's formula, $H_d = \mathrm{Tr}\,(\mathbf{C_p}) - (\mathbf{1'}\,\mathbf{C_p}\,\mathbf{1})/m$. The equivalent of identifying the variable with the smallest regression weights after minimizing these weights by subtracting the appropriate constant (i.e., the mean value), is obtained by retaining the variables in $\mathbf{C_p}$ which provide the largest value of $H_d$ as a function of $\mathbf{C_p}$.

The sums of the squared elements of each row of $\mathbf{F_{t1}}$ are equal to 1.0. This sum of squares for the remaining rows, for all rows below $\mathbf{F_{t1}}$, is equal to $(R_{i(1,2,3)})^2$, the multiple correlation coefficient between the $i^{th}$ variable and the least square prediction of the $i^{th}$ variable based on the first three selected variables.

An algorithm for creating triangular solutions like $\mathbf{F_{t1}}$ bordered below by $\mathbf{F_{t2}}$ (e.g., the "square root" algorithm ) as shown above is readily extended to additional variables for which the correlation coefficients with variables 1, 2, and 3 are known. The $\mathbf{F_v}$ created by this triangular factorization algorithm is a factor extension solution; thus $\mathbf{F_v}$ equals $\mathbf{V}\,(\mathbf{F_{t1}}'\,\mathbf{F_{t1}})^{-1}\,\mathbf{F_{t1}}$ in accordance with Dwyer's formula for provision of a factor extension solution.

The square root factorization algorithm uses a transformation matrix comparable to a $\mathbf{T}$ matrix described in Appendices 2A and 2B as the multiplier of the $k$ orthogonal factors bordered by an oblique factor to create the $(k+1)^{th}$ orthogonal factor. The oblique factor is the $(k+1)^{th}$ predictor selected for inclusion in the orthogonal factor solution. The $\mathbf{T_{k+1}}$ that creates the solution for an additional orthogonal factor to be added to $\mathbf{F_{2t}}$ provides loadings on the same new factor for the criterion variable. This successive adding of factors to the test space and the extending of these factors to the criterion space is described in Appendix 3C.

Horst's "accretion" algorithm for successively selecting tests to maximize his "absolute validity" index, $H_a$, calls for producing a triangular factor solution and extending this solution to the other variables. The rows corresponding to the criterion variables define a matrix equivalent to the factor extension matrix, $\mathbf{F_v}$. Using the same notation as above we will examine an example with three jobs, e, f, and g, with respect to the first three tests selected by accretion. $H_a$ is constructed as the sum of the squared elements of each column of $\mathbf{F_v}$, referred to as: $H_{a1}$, $H_{a2}$, $H_{a3}$. For our example we define each of these sums of squares, $H_{aj}$, as follows:

$H_{aj} = (r_{e1})^2 + (r_{f1})^2 + (r_{g1})^2$, for j = 1;

$H_{aj} = (r_{e(2.1)})^2 + (r_{f(2.1)})^2 + (r_{g(2.1)})^2$, for j = 2;

$H_{aj} = (r_{e(3.12)})^2 + (r_{f(3.12)})^2 + (r_{g(3.12)})^2$, for j = 3.

In the accretion test selection procedure the value for $H_{aj}$ is successively maximized through the judicious selection of the next test, keeping all previous selected tests;

$$H_a = \mathrm{SUM}_j^k\, H_{aj}.$$

The "accretion" test selection sequence in which $H_a$ is successively computed is comparable to the formula $H_a = \mathrm{tr}\,(\mathbf{F_v'F_v})$. Horst's index of absolute validity can also be written as a sum of the squared multiple correlation coefficients. For the above example this would be $H_a = \mathrm{SUM}_i^m\,(R_{i(123)})^2$, or in matrix notation, $H_a = \mathrm{tr}\,(F_v\,F_v')$. In this example, $H_a = (R_{e(123)})^2 + (R_{f(123)})^2 + (R_{g(123)})^2$ and $(R_{i(123)})^2 = (R_{i1})^2 + (R_{i(2.1)})^2 + (R_{i(3.12)})^2$, for i equal to e, f, and g.

Similarly, Horst's "differential validity" index, $H_d$, can be defined in terms of successively determined values of $H_{dj}$ where each such value, as with $H_{aj}$, represents the contribution of the orthogonal components of the selected variables (tests or factors) to the overall index $H_d$. In our above example $H_d = \mathrm{SUM}_j^k\,H_{dj}$, where $H_{dj} = (r_{e1} - r_1^*)^2 + (r_{f1} - r_1^*)^2 + (r_{g1} - r^*)^2$, for j = 1; $H_{dj} = (r_{e(2.1)} - r_2^*)^2 + (r_{f(2.1)} - r_2^*)^2 + (r_{g(2.1)} - r_2^*)^2$, for j = 2; $H_{dj} = (r_{e(3.12)} - r_3^*)^2 + (r_{f(3.12)} - r_3^*)^2 + (r_{g(3.12)} - r_3^*)^2$, for j = 3; $r_j^*$ is the mean of the $j^{th}$ column of $F_v$. In matrix notation, $H_d = \mathrm{tr}\,[(F_v - HF_v)'(F_v - HF_v)]$, where $\mathbf{H}$ equals an $m$ by $m$ matrix whose elements all equal $(1/m)$.

As true with respect to $H_a$, $H_d$ is successively maximized in Horst's accretion algorithm through the judicious selection of the next test to be added to the test battery. $H_d$ can also be computed in terms of an orthogonal rotation of $\mathbf{F_v}$, i.e., $\mathbf{F_v\,A}$, as follows: $H_d = \mathbf{(F_v\,A - H\,F_v\,A)\,(F_vA - HF_vA)'}$. When rewritten as $H_d = \mathbf{(F_v - HF_v)\,AA'\,(F_v - H\,F_v)'}$,it becomes obvious that any transformation matrix such that $\mathbf{AA' = I}$, as would be true of any orthogonal transformation, will give the same value for $H_d$. Thus while Horst made use of an $\mathbf{F_v}$ which was a triangular factor solution, any other orthogonal transformation of $\mathbf{F_v}$, that is any factor extension of $\mathbf{F_t}$, or orthogonal transformation of $\mathbf{F_t}$, would serve just as well.

# APPENDIX 2D
# HORST'S CONCEPT OF DIFFERENTIAL VALIDITY AND HIS $H_D$ INDEX

It seems to us from Horst's (1954) description of his differential validity (DV) index that he first provided a basic concept formulating a measure he intuitively believed to be related to classification efficiency, and then provided computational simplifications for the case where the interest is in PCE, rather than in CE. Since we believe DV has a more general importance than the provision of an intuitive basis for $H_d$, we first present a more detailed description of the basic concept of DV and then carefully point out the manner in which Horst chose to restrict his DV concept in creating his simplified formulae for $H_d$.

The general concept of DV can be described as the prediction of the criterion differences, between pairs of PP scores, by the predictor differences between corresponding pairs of predictor variables. Intuitively an efficient classification process implies being able to decide effectively between each pair of possible assignment alternatives. The overall index of decision effectiveness is the aggregate of all of these pairwise decisions. The DV index measures the covariance between the predictor difference scores and the criterion difference scores.

Horst's DV index is a measure of classification efficiency obtainable when the maximally effective assignment variables (FLS composites) are used. We have consistently referred to this kind of efficiency as potential classification efficiency (PCE) as contrasted to a measure of the classification efficiency of the operational composites that are not FLS composites; the latter is simply classification efficiency or CE. Since measurement of the PCE of the battery requires the use of FLS composites (i.e., predicted performance or PP variables) as the basis of the predictor differences and these same variables are appropriately used as the surrogate criterion variables, it becomes possible to greatly simplify the computing formula. This simplified computing formulae must not be used when the assignment variables are not FLS composites; the application of Horst's DV concept to measure the CE of operational ASVAB aptitude areas, or any other set of test composites that are not FLS composites, must commence with a more basic formulation.

Conceptually, the unit of analysis for computing a DV index for measuring PCE is each possible pair of criterion scores in the sample: the $j^{th}$ criterion paired with the $k^{th}$ criterion for the $i^{th}$ individual. The classification decision is visualized as one of distinguishing between the $j^{th}$ and $k^{th}$ job for $m(m-1)$ pairs of different jobs. Each pair of jobs is matched with a corresponding pair of predictor variables. It is readily seen that Horst's DV index of PCE ($H_d$) is almost, but not quite, a correlation coefficient; it is actually a covariance. Before simplification this index is a sum of $m(m-1)$ cross products of difference scores divided by $m$, where $m$ is the number of jobs.

The difference between the $j^{th}$ and $k^{th}$ criterion score will be denoted as $d_{jk}$ and corresponding difference for the predictor scores as $p_{jk}$. Each cross product, $c_{jk}$, is equal to $d_{jk}$ times $p_{jk}$. Thus a cross product, $c_{jk}$, one for each unit of analysis,is as follows: $c_{jki} = (y_{ji} - y_{ki})\,(z_{ji} - z_{ki})$.

We assign the symbol $H_p$ to the more general concept of the DV index which does not assume the equality of $p_{jk}$ and $d_{jk}$. Horst did not discuss the possibility of using predictor pairs other than FLS composites; we restrict our more general model to predictors that are PP variables (but not necessarily FLS estimates) with standard deviations equal to their validities. Using ($d_{jk}$) to designate an $N$ by $m^2$ matrix of criterion difference scores and ($p_{jk}$) to designate the corresponding $N$ by $m^2$ matrix of predictor difference scores we can write $H_p$ as follows: $H_p = (1/m)$ tr $((pjk)'(djk)$. Each of the $m^2$ diagonal elements of $(pjk)'(djk)$ takes the general form: $\mathrm{SUM}_i^N\ (y_j - y_k)(z_j - z_k) = y_j z_j + y_k z_k - 2y_j z_k$. As $j$ and $k$ each take all values from 1 to $m$ the sum of these 4 $m^2$ terms can be written in terms of the score matrices **Y** and **Z** as follows:

$2\,(\mathrm{tr}\ \mathbf{Y}'\,\mathbf{Z}) - \mathbf{1}'\,(\mathbf{Y}'\mathbf{Z})\,\mathbf{1})$. Since both $H_d$ and $H_p$ are based on only the $m(m-1)/2$ different pairs, and m of the $m^2$ difference scores for each individual are equal to zero, $H_p = \mathrm{tr}\ (\mathbf{Y}'\mathbf{Z}) - (\mathbf{1}'(\mathbf{Y}'\mathbf{Z})\,\mathbf{1})/m$. Using similar logic, $H_d = \mathrm{tr}\ (\mathbf{Z}'\mathbf{Z}) - (\mathbf{1}'(\mathbf{Z}'\mathbf{Z})\ \mathbf{1})/m = \mathrm{tr}\ (\mathbf{C}_p) - (\mathbf{1}'\mathbf{C}_p\,\mathbf{1})/m$.

Just as $\mathbf{F}_t$ can be extended to the $m$ criterion variables yielding the extended factor solution, $\mathbf{F}_v$, this solution can be extended, using the same approach, to the $m(m-1)$ variables defined as the differences between the $j^{th}$ and $k^{th}$ criterion variables. The factor loadings of these $m(m-1)$ difference variables on the factors found in $\mathbf{F}_t$ and $\mathbf{F}_v$ provide a factor solution we refer to as $\mathbf{F}_h$.

The criterion differences can be expressed in terms of PP scores, as in $(z_j - z_k)$ or as in the differences between the rows of $\mathbf{F_v}$. The rows of $\mathbf{F_h}$ can be duplicated as the differences between the rows of $\mathbf{F_v}$. Thus, we can define $H_d$ in terms of either $\mathbf{F_h}$ or $\mathbf{F_v}$.

We now consider a set of matrices, one for each individual, whose general term is $(z_j - z_k)^2$. This matrix, $\mathbf{M_i}$, is defined as follows:

$$\mathbf{Mi} = \begin{bmatrix} (z_1 - z_1)^2, & (z_1 - z_2)^2, & \ldots\ldots\ldots & (z_1 - z_m)^2 \\ (z_2 - z_1)^2, & (z_2 - z_2)^2, & \ldots\ldots\ldots & (z_2 - z_m)^2 \\ \ldots\ldots\ldots, & \ldots\ldots\ldots, & \ldots\ldots\ldots, & \ldots\ldots\ldots \\ (z_m - z_1)^2, & (z_m - z_2)^2, & \ldots\ldots\ldots & (z_m - z_m)^2 \end{bmatrix}.$$

Note that the sum of the elements of all $\mathbf{M_i}$, each individual's matrix, across $N$ individuals, (i.e., $\mathrm{SUM}_i^N\ M_i$) divided by $N$, equals double the value of $H_d$. Considering the $j^{th}$ column of each $\mathbf{M_i}$ separately, we see that summing across all $N$ matrices and dividing by $N$ yields the squared standard deviation of the PP variables ($S_z{}^2$) around the grand mean of these criterion variables plus the squared difference between the mean of the jth criterion score and the grand mean. Summing over the $m$ columns and dividing by $N$ times $m^2$, the total number of terms, yields two times $S_z{}^2$. Since the m diagonal terms of each $\mathbf{M_i}$ have zero values, the average of the $m(m-1)$ terms either above or below the diagonal yield a value of $S_z{}^2$. We see that $H_d$ equals $m$ times $S_z{}^2$ since Horst divided his sum of squares by $N$ times $m$, rather than $N$ times $m^2$.

As shown in a paragraph above, $S_z{}^2$ can also be expressed in terms of $\mathbf{F_v}$. A vector in which each element is the mean of the corresponding column of $\mathbf{F_v}$ represents the mean of the PP scores in terms of each of the regression weights of each of the column variables. The sum of the squared deviations of each row vector of $\mathbf{F_v}$ around this "mean" vector gives the product of m and $S_z{}^2$ and is thus seen to be equal to $H_d$. We find it convenient to write $m(S_z)^2$ in matrix notation as follows: $m(S_z)^2 = H_d = \mathrm{tr}\ [(\mathbf{F_v} - \mathbf{HF_v})(\mathbf{F_v} - \mathbf{HF_v})']$ where $\mathbf{H}$ is an m by m matrix for which every element is equal to 1/m. Each element of every column of $\mathbf{HF_v}$ is equal to the mean of that column so that $(\mathbf{F_v} - \mathbf{HF_v})$, referred to as the matrix $\mathbf{G}$ elsewhere, is a matrix for which the sum of its squared elements is equal to $H_d$.

We demonstrated above that $H_d = \mathrm{tr}\ (\mathbf{C_p}) - \mathbf{1'}\ \mathbf{C_p}\ \mathbf{1}\ (1/m)$. Horst (1954) provides this formula in different notation on page 25, formula 43. It is very easy to demonstrate that tr $(\mathbf{F_v} - \mathbf{HF_v})(\mathbf{F_v} - \mathbf{HF_v})'$is equal to tr $(\mathbf{C_p}) - \mathbf{1}'\, \mathbf{C_p}\, \mathbf{1}\, (1/m)$. Multiplying out $\mathbf{GG}'$, we have tr $(\mathbf{F_v F_v}')$ + tr $(\mathbf{HF_v F_v' H'})$ – tr $(\mathbf{HF_v F_v}')$ – tr $(\mathbf{F_v F_v' H'})$. When $\mathbf{F_v}$ is a complete factorization of $\mathbf{C_p}$, $\mathbf{F_v}\, \mathbf{F_v}' = \mathbf{C_p}$; this covariance matrix is included in each of the four terms. The first of these four terms is equal to tr $(\mathbf{C_p})$ and we only need to show that the remaining three terms are equal to $-(\mathbf{1'C_p1})/m$ to complete our demonstration.

It is obvious from its definition that $\mathbf{H} = \mathbf{H}'$, and only a little less obvious that $\mathbf{HC_p} = \mathbf{C_pH}'$ since, for a symmetrical matrix such as $\mathbf{C_p}$, the means of the columns must equal the means of the corresponding rows. The remaining three terms can be written as follows:

| | | |
|---|---|---|
| tr $\mathbf{HC_p}$ | $=$ SUM$_j^m$ (mean of the $j^{th}$ column of $\mathbf{C_p}$) | $= (\mathbf{1'C_p1})1/m$ |
| tr $\mathbf{C_pH}$ | $=$ SUM$_j^m$ (mean of the $j^{th}$ row of $\mathbf{C_p})_j$ | $= (\mathbf{1'C_p1})1/m$ |
| tr $\mathbf{HC_pH}$ | $= m^2$ times [(grand mean of all elements of $\mathbf{C_p})/m$] | $= (\mathbf{1'C_p1})1/m$ |

Thus, the sum of the remaining three terms, considering signs, is seen to be minus $(\mathbf{1'C_p1})/m$ and the equality of tr $(\mathbf{GG}')$ to tr $(\mathbf{C_p})$ minus $(\mathbf{1'C_p\, 1})/m$ is proven.

# APPENDIX 2E
# APPLICATION OF A DIFFERENTIAL VALIDITY CONCEPT TO MEASURE THE CLASSIFICATION EFFICIENCY OF SETS OF TEST COMPOSITES

The index $H_d$ provides an approximate measure of the potential classification efficiency (PCE) of a predictor battery. This index is of no use when it is desired to measure the classification efficiency (CE) of an existing set of operational test composites that are not full least square (FLS) estimates of performance on the jobs for which they are used as assignment variables. The value of PCE for a battery provides an upper bound for the CE that is obtainable for any set of test composites.

We believe that the index, $H_p$, described in the previous appendix is as appropriate for measuring the CE of a set of operational assignment variables as $H_d$ is for determining the PCE of a test battery. In this appendix we describe a practical approach for using this index as an approximate measure of CE.

Using the notation of Appendix 2A, an $N$ by $m$ matrix of predictor scores, in standard score form--with each element divided by the square root of $N$--is written in underlined bold face type as $\underline{\mathbf{Y}}$. An $N$ by $m$ matrix of FLS estimates of the performance measures of $m$ jobs, also with each element in standard score form and divided by the square root of $N$, is written as $\underline{\mathbf{Z}}$. Thus $\mathbf{S}_p^{-1/2}\underline{\mathbf{Z}}'\underline{\mathbf{Y}} = \mathbf{V}$, where $\mathbf{V}$ is an $m$ by $m$ matrix of validity coefficients whose rows represent the jobs and the columns the corresponding predictor composites.

To convert $\mathbf{V}$ into the covariance matrix of interest, we need two $m$ by $m$ diagonal matrices whose non-zero elements are SDs: $\mathbf{S}_y$ has the same diagonal elements as does $\mathbf{V}$, with zeros elsewhere; $\mathbf{S}_p$ has as its diagonal elements the validities of each FLS estimate of job performance, and zeros elsewhere. The order of the $y$ and $z$ variables in these two diagonal matrices must, of course, correspond. Using these three values we can define $H_p$ as follows:

$$\mathbf{C}_{zy} = \mathbf{S}_p \mathbf{V} \mathbf{S}_y, \tag{1}$$

$$H_p = \mathrm{tr}\,(\mathbf{C}_{zy}) - \mathbf{1}'\,\mathbf{C}_{zy}\,\mathbf{1}(1/m) \tag{2}$$

When the predictor composites in the **Y** matrix are FLS estimates, our $H_p$ index becomes $H_d$ and is an estimate of both CE and PCE.

MacLaughlin et al. (1984) proposed dividing what we refer to as $H_d$ by $m$ to obtain the index they call $H^2$; $H_d/m = H^2$. They proposed using their own index, $M$, for measuring the CE of alternative operational sets of test composites. $M$ was then divided by $H$ to estimate the ratio of CE/PCE. While we would not use $M$, we like their proposed use of CE/PCE as a means of estimating the relative classification efficiency of alternative sets of classification composites and/or alternative job families.

The comparison of $H_p$ indices for sets of composites possessing different values of $m$ poses a difficult problem. The solution of McLaughlin et al. was to divide their raw indices by $m$, a process comparable to the division of $H_p$ and $H_d$ by $m$ to obtain a value that is the actual covariance, rather than m times the covariance. Unfortunately this covariance value does not reflect the greater PCE and CE that comes with a larger value of $m$. $H_d$ and $H_p$ undivided by $m$ greatly overestimate the increase in PCE or CE that results from increasing $m$. On the other hand, dividing $H_d$ or $H_p$ by $m$ creates almost as large an underestimate in some ranges of $m$. Dividing by $m$ is definitely not the answer to the problem, unless a further compensating correction is made.

We propose using the multipliers provided by Brogden (1959) to reflect the effect the number of test composites and associated job families have on PCE. This multiplier is referred to in the text as $M_{pm}$, with p standing for the percent rejected and m for the number of jobs; The symbol $B_m$ is used here to stand for $M_{0m}$. We propose using these multipliers on both PCE and CE estimates. The corrected $H_p$ and $H_d$ indices to be used for comparing sets of composites of differing $m$ would then be as follows:

$$H_{dc} = ((H_d)^{1/2}/m)B_m \quad , \tag{3}$$

$$H_{pc} = ((H_p)^{1/2}/m)B_m \quad , \tag{4}$$

where $B_m$ takes on the values from Brogden's table that shows: $B_2 = 0.56$; $B_3 = 0.85$; $B_4 = 1.03$; $B_5 = 1.16$; $B_6 = 1.27$; $B_7 = 1.35$; $B_8 = 1.42$; $B_9 = 1.49$; $B_{10} = 1.54$; $B_{11} = 1.59$; $B_{12} = 1.63$; $B_{13} = 1.67$; $B_{14} = 1.70$; $B_{15} = 1.73$. For $m$ greater than 15, non-linear extrapolation should provide adequately accurate values for $B_m$. The ratio of CE/ would then be computed as $H_{pc}/H_{dc}$.

The value of $m$ used in computing $H_{dc}$ and $H_{pc}$ will usually be different. McLaughlin et al. based their value of H on the total number of jobs for which they had validities ($m = 98$) while their alternative sets of composites were all less than the number in the current operational battery ($m = 9$). We would similarly base $H_d$ on as many FLS composites as the data will permit the computation of moderately stable FLS regression weights. The use of $m$ and $B_m$ in conjunction with *the square root of* $H_d$ appears to be justified by the relationship between $H_d$ and PCE indicated in Appendix 2G.

# APPENDIX 2F
# COMPARISON OF BROGDEN'S CLASSIFICATION EFFICIENCY MEASURE WITH HORST'S DV INDEX, $H_d$

Brogden's 1959 model is based on a set of assumptions regarding the relationships among and across predictor and criterion variables, relationships that can be depicted in terms of Spearman's Two Factor theory. These assumptions are met if the factor matrix $\mathbf{F}_v$, a matrix such that $\mathbf{F}_v\mathbf{F}_v'$ is equal to $\mathbf{C}_p$, all elements of the first (general) factor are equal to the product $R(r)^{1/2}$ and the remaining $m$ (job specific) factors can be expressed as a diagonal matrix with the diagonal elements equal to $R(1-r)^{1/2}$. A three job example would appear as follows:

$$\mathbf{F}v = \begin{bmatrix} R(\mathrm{r})^{1/2} & R(1-r)^{1/2} & 0.0 & 0.0 \\ R(\mathrm{r})^{1/2} & 0.0 & R(1-r)^{1/2} & 0.0 \\ R(\mathrm{r})^{1/2} & 0.0 & 0.0 & R(1-\mathrm{r})^{1/2} \end{bmatrix}.$$

As described in Appendix 2A, $r$ represents the common correlation coefficient among the predictor composites (all FLS estimates), and $R$ is the common multiple correlation coefficient (the validity coefficient) of these predictor composites. $\mathbf{C}_p$ is the matrix of covariances among the FLS estimates of job performance and thus would have all diagonal elements equal to $R^2$ and all off-diagonal elements equal to $(R^2)r$. It is readily seen that the matrix of correlation coefficients among the FLS estimates is equal to $\mathbf{S}_p{}^{1/2}$ $\mathbf{C}_p$ $\mathbf{S}_p{}^{1/2}$, where $\mathbf{S}_p$ is a diagonal matrix whose diagonal elements are equal to the diagonal elements of $\mathbf{C}_p$. The diagonal elements of this correlation matrix are ones and the off-diagonal elements are all equal to $r$.

Horst's $H_d$ index is equal to the sum of the squared deviations from the column means of each element of $\mathbf{F}_v$. Looking at $\mathbf{F}_v$ as defined to fulfill Brogden's assumptions, we see that the sum of squared deviations for the first column of $\mathbf{F}_v$ is zero and is $R(1-r)$ for each of the other m columns. Thus $H_d$ is equal to (m − 1) times $R(1-r)$ when Brogden's assumptions are met. The same result is obtained if the $\mathbf{C}_p$ described above is entered into the formula: $H_d = \mathrm{tr}\,(\mathbf{C}_p) - \mathbf{1}'\,\mathbf{C}_p\,\mathbf{1}(1/m)$. Since PCE is defined by Brogden as equal to $M_{mp}$ times $R(1 - r)^{1/2}$, we see that we need to take the square root of $H_d$, divide by (m − 1), and multiply by $M_{pm}$ (a value tabled by Brogden) to obtain PCE when Brogden's assumptions are met.

# CHAPTER 3. IMPROVING CLASSIFICATION EFFECTIVENESS

## A. HORST'S TEST SELECTION APPROACHES

There is no procedure that can make a set of weighted test composites of a fixed battery more effective for use as classification tools than the weighted composites optimized for selection (i.e., the LSEs). For both objectives, the optimal weights to use are the least squares weights and the maximally effective composite is a LSE. However, it is possible to reduce potential classification efficiency (PCE) more than necessary in the process of creating test composites that have fewer tests than contained in the operational test battery, or in selecting weights for the tests in a composite that are other than the least squares weights for predicting the criterion. There are useful techniques for minimizing the loss in PCE in selecting test composites from an operational battery as well as procedures for maximizing PCE of an operational battery selected from a larger experimental battery.

When a subset of tests is to be selected from a larger set of experimental tests for use as an operational battery, a subset selected to maximize classification efficiency will have more PCE/PAE than a set selected to maximize selection efficiency. Harris (1967)through simulation showed that subsets selected to maximize Horst's index of differential validity, $H_d$ were superior to subsets selected to maximize Horst's index of absolute validity, $H_a$ -- the sum of the squared multiple correlation coefficients across the jobs included in the test selection and assignment simulation. Harris' simulation methodology and Horst's indices are detailed in Chapter 2.

Horst's differential and absolute techniques for providing test selection against multiple criteria (Horst, 1954, 1955) can be compared to a common accretion approach in sequentially selecting tests against a single criterion. A sequential method first selects the test with the highest validity for inclusion in the battery; next, each of the remaining tests is paired with the selected test and the pair yielding the highest multiple correlation coefficient is considered "best" and retained, and then these two are matched with each of the remaining tests and the "best" triad of tests that includes the first two selected tests is retained. This process is continued until the desired number of tests is selected or no remaining test can make a practical contribution to the magnitude of the multiple correlation coefficient.

Horst's "absolute" method differs from the usual approach that focuses on a single criterion in that it utilizes as an index ($H_a$) the sum of the squared multiple correlation coefficients across some specified number ($m$) of jobs. This index, $H_a$, is maximized at each step. Similarly, Horst's "differential" method maximizes $H_d$ at each step.

The extended factor matrix, **F**, described in the previous chapter as a Dwyer factor extension matrix, is obtained by extending a complete factorization of the intercorrelations among the predictor tests, $\mathbf{R}_t$, into the criterion space. Although by no means apparent from Horst's presentation (1954), **F** is in effect constructed, column by column, by Horst's test selection process. In this test selection procedure the implied **F** matrix is a factor extension of a triangular factorization of $\mathbf{R}_t$; it contrasts to the more general representation of **F** as the Dwyer factor extension of any complete factorization of $\mathbf{R}_t$ (with ones in the diagonals).[14]

The first column of **F** consists of the correlation coefficients of the $i^{th}$ criterion variable with the first test selected, that is, a column vector of values for $r_{i1}$. The second column of **F** will be the semipartial correlation coefficients between each criterion variable and the component of the next selected test that is uncorrelated (orthogonal to) the first selected test, that is, a column vector of values for $r_{i(2.1)}$. The first test is selected because its use maximizes either $H_a$ or $H_d$, depending on which is to be used as the figure of merit. $H_a$ at that stage is measured as the sum of the squared values of $r_{1i}$, and $H_d$ is measured as m times the variance of each trial column vector, ($r_{i1}$). The variable to be designated as "1" is, of course, designated as such only after every test in that role has been tried out.

Similarly, each of the remaining tests is tried out to see which one will maximize the sums of its squares ($H_a$), or $m$ times the variances of the semipartial correlation coefficients ($H_d$), in the second column of **F**. As additional tests are selected, the $i^{th}$ row of **F** can be depicted in terms of semipartial correlation coefficients as follows:

$$\mathbf{F} = (r_{i1}\ ,\ r_{i(2.1)}\ ,\ r_{i(3.12)}\ ,\ r_{i(4.123)}\ ,\ \ldots\text{etc.})$$

14 Horst (1954) does not mention the triangular factorization of $\mathbf{R}_t$ nor the extension of this solution into the criterion space. Instead he cites one of Dwyer's algorithms that is efficient for hand computations but adds little to the understanding of the process.

The sums of squares of the elements of each of these rows is obviously the squared multiple correlation coefficient between the $i^{th}$ criterion variable and the LSE. It is less obvious, but equally true, that the squared differences from the column means of **F** summed for a row indicates the contribution of a job to the total differential validity with respect to its pairing with each of the other jobs.

We consistently use $\mathbf{F}_t$ as the factor matrix that reproduces $\mathbf{R}_t$, (i.e., $\mathbf{F}_t \mathbf{F}_t' = \mathbf{R}_t$). In our description of Horst's test selection procedure, $\mathbf{F}_t$ denotes a square root (triangular) factor matrix. The rows in $\mathbf{F}_t$ have the same type of semipartial correlation coefficients as the rows in **F**; the $i^{th}$ variable is a test instead of a LSE or a job criterion, and the sums of squares of the row elements are unity instead of the squared multiple correlation coefficients found in **F**. Each column of $\mathbf{F}_t$, after the first, is the correlation of a test with a component of the selected variable that is orthogonal to the variables represented by all columns to the left. These column variables are assigned a variance of one and are mutually orthogonal, and thus can be considered as factors. The column variables in **F** are exactly the same variables, column by column, as the column variables of $\mathbf{F}_t$.

The number of columns in **F** grows by one as each additional variable is selected. At each stage in the test selection process, depending on whether $H_a$ or $H_d$ is being maximized, one of the following relationships holds[15]:

$$H_a = \text{tr}\,(\mathbf{FF'}) = \text{tr}\,(\mathbf{F'F}) \quad . \tag{3.1a}$$

$$H_d = \text{tr}\,(\mathbf{F-HF})\,(\mathbf{F-HF})' = \text{tr}\,(\mathbf{F-HF})'\,(\mathbf{F-HF}) \quad . \tag{3.1b}$$

The **F** built up by either of the test selection processes, "absolute" or "differential," will have as many columns as there are selected tests. In either case, $\mathbf{FF'} = \mathbf{C}$, and $\mathbf{FF}_t' = \mathbf{V}$, where **V** is the validity matrix for the selected tests. It should be noted that the particular **F** used in this test selection process is only an orthogonal rotation away from any other **F** that meets the more general definition mentioned above. Any orthogonal rotation of a Dwyer factor extension of any complete factorization of $\mathbf{R_t}$, that is, a Dwyer factor extension of any alternative $\mathbf{F}_t$, fulfills the more general definition of **F**.

In Horst's test selection procedure this **F** is by implication directly created by performing the same operations on **V** that are performed on $\mathbf{R_t}$ to create $\mathbf{F_t}$ as a triangular matrix. A more general solution of **F** in terms of the validity matrix, **V**, and any $\mathbf{R_t}$ is as follows: $\mathbf{F} = \mathbf{VF_t'(F_t'F_t)^{-1}}$ (Dwyer, 1937). This same Dwyer factor extension formula

---

15 See previous chapter for explanation of the matrix **H** and the operator **tr**.

is, of course, applicable when $\mathbf{F}_t$ is a triangular factorization of $\mathbf{R}_t$. Thus the indices $H_a$ and $H_d$ as used in the test selection process still fall within the scope of the development and discussion of $H_a$ and $H_d$ provided in the previous chapter.

Since a sequential selection of tests does not guarantee that the selected set of tests provides the maximum possible value for $H_d$ (or $H_a$), it may be desirable to try out other sets of tests to see if any alternative set would yield higher index values. For example, if tests were chosen from a larger pool of tests, and only 9 tests are desired in the battery, it would be reasonable to compute $H_d$ or $H_a$ for a test set in which the tenth selected test was substituted for the 9th selected. Also it may be desirable to compare a test set selected by a different process with one selected by the Horst sequential method. The Horst indexes can be readily computed from any factor extension matrix $\mathbf{F}$, or from any orthogonal rotation of $\mathbf{F}$ (see Appendix 3C).

It was demonstrated in the previous chapter that the presence of hierarchical classification effects could cause $H_d$ to lose its proportionality to the square of PCE. Adjusting the rows of $\mathbf{F}$ to make each job vector the same length should prevent the $H_d$ value from being affected by unequal validities. To incorporate this adjustment into the test selection process, an adjustment is made on each semipartial correlation coefficient before it is subtracted from the column mean and the difference squared. This adjustment can be accomplished by the following formula to provide an index that can be used as the figure of merit in the sequential test selection procedure:

$$\text{Adjusted } H_d = \Sigma_j^k \left( \Sigma_i^m \left( (a_{ij})\, \overline{R}/R_i - \bar{a}_j \right)^2 \right) \quad ; \tag{3.2}$$

The figure of merit used for the test selection process is the inner sum, the value for one column; where $\bar{a}_j = \Sigma_i^m a_{ij}$ , $\mathbf{F} = [a_{ij}]$ , and $\Sigma_j^k (a_{ij})^2 = (R_i)^2$ .

Job samples on which data have been collected vary in size but, more important, expected job quotas used in the future assignment process are far from equal. Similarly, the MPP standard score, PCE, that results from making optimal assignments to fill quotas over a period of time is influenced by the differential validity attached to jobs with the larger quotas. The contribution of each row of $\mathbf{F}$ to $H_d$ is due to the differential potential of the comparisons of one job, the job represented by that row, with each of the other jobs. The importance of each row is thus proportional to the expected quota of the associated job. The formula for $H_d$ weighted by the quota weight for the $i^{\text{th}}$ job ($W_i$) is as follows:

$$\text{Weighted } H_d = \Sigma_j^k \left( \Sigma_i^m W_i (a_{ij} - \bar{a}_j)^2 \right) \quad . \tag{3.3}$$

The figure of merit used in the test selection process is the inner sum, the value for one column.

## B. THE COMPARISON OF HORST'S AND BROGDEN'S APPROACHES TO TEST SELECTION FOR THE IMPROVEMENT OF CLASSIFICATION EFFICIENCY

Rather than sequentially selecting tests from an experimental test pool for retention in an operational battery, several authors have proposed methods for sequentially eliminating the least effective tests. Horst and MacEwan (1960) have provided such a sequential elimination process for arriving at a subset of tests that provide close to the maximum value for $H_d$. The set selected by such a deletion type sequential test selection process can be compared with a set selected by the accretion type sequential test selection process described in the previous section. If the two sets identified by the two methods are identical, one can safely assume that $H_d$ is truly maximized; it is highly unlikely that there is any other set with the same number of tests that yields a higher $H_d$. However, if the $n$ selected tests have $k$ tests in each set that are not common to both sets, one could assume that $n$-$k$ tests, the overlapping ones, can be safely adopted, while all combinations of the remaining $2k$ tests should be considered, in every combination of k tests, as possible members of the set that truly maximizes $H_d$.

The Horst and MacEwan elimination method can be visualized as being equivalent to, in terms of results, the procedure obtained by: (1) computing the matrix of covariances among LSEs, that is, $\mathbf{C} = \mathbf{V}(\mathbf{R}_t)^{-1}\mathbf{V}'$, and then computing either $H_d$ or $H_a$; (2) computing $\mathbf{C}$ with each test removed, in turn, from the test pool (for $n$ tests in the pool, $n$ different $\mathbf{C}$ matrices and a $H_d$ or $H_a$ for each $\mathbf{C}$ is computed); (3) identifying the test to be deleted by its absence from the set used in the computation of the $\mathbf{C}$ yielding the largest $H_d$ or $H_a$, and permanently removing this test from the pool, and, finally, (4) repeating all steps until the elimination of further tests would reduce the index by an unacceptable amount, or the desired number of tests for the operational battery has been achieved.

An alternative approach for accomplishing the same results as the Horst and MacEwan deletion procedure is one that is less computationally efficient, but highlights a point of similarity with an approach propo... by Brogden (Brogden 1964). In this approach an $\mathbf{F}$ matrix is produced as the factor extension of a full triangular factorization of

$\mathbf{R_t}$ and **F** is recomputed with each trial column in turn, placed in the last (right most) position. The last column shows the contribution of a trial variable (being considered for deletion) to the criteria when the effects of all other variables have been removed. Eliminating the predictor variable whose semipartial correlation coefficients (when located in the last column) have the smallest variance, is equivalent to the elimination of the test for which the regression weights (in this case for orthogonal components) have the smallest variance; each of these elements of **F** is a regression weight being applied to the test component that is uncorrelated (orthogonal to) all other tests. The smallest variance may result because the regression weights are either very small, or because they are similar to each other across jobs. Intuitively, this particular test elimination procedure appears to eliminate systematically the same tests as the method proposed by Brogden.

Brogden (1964) pointed out that the regression weights in LSEs can be directly examined to identify tests that are making no contribution to classification efficiency. Such a weight matrix could be designated as **W** and expressed in our notation as follows: $\mathbf{W} = \mathbf{R}_t^{-1}\,\mathbf{V}'$.

Noting that classification efficiency is not affected if constants are added to the columns of **W**, Brogden proposed that constants be judiciously added so as to, hopefully, provide, for some test, a column array of regression weights with *zero*, or *near zero*, values. A test with zero weights in all composites could obviously be dropped from the battery. Such a process is directed at the deletion rather than the accretion of tests selected to form a battery.

Once a test has been removed from the battery by the method just described, the **W** matrix can be recomputed and column constants again judiciously added in search of another test that can be deleted. This process could be repeated until no further candidates for removal can be identified.

The Brogden deletion method would be useful for removing tests from some LSEs, but not from others, in order to make test composites (aptitude areas) of more manageable size, while at the same time minimizing the decrease of classification efficiency. Also this approach could be used to eliminate negative weights in test composites based on LSEs without reducing classification effectiveness.

It is clear that the addition of column constants to **F**, also a matrix of regression weights for application against variables (factors) for the prediction of either the criteria or the LSEs, has no effect on the magnitude of either PCE or $H_d$. Also, the columns of **F** can be adjusted by the addition of a constant to the weights of a given test component (or factor in the more general case) across all rows of **F** to eliminate negative weights.

The use of a LSE associated with each job for use in the assignment process, combined with a smaller set of test composites to be used first by a counselor/classifier, and then recorded on the official record for later operational use, is a distinct practical possibility. The recorded composite scores would be available for operational use, (including by the examinee) to determine minimum eligibility for military programs or training. The use of a small number of composite scores results in a high probability that the prediction of success in a larger number of specific jobs would require weighting of the composites (the computing of LSEs, where the composites are the independent variables and the job criteria are the dependent variables). Some weights in the composites may be negative: there are motivational and administrative limitations on the visible use of negative weights in producing scores resulting in important personnel decisions. Adding constants to eliminate negative weights, without affecting classification efficiency, is one attractive option. The practical use of providing a relatively small set of scores to the counselor making classification decisions in the military setting is discussed further in a later section.

## C. AN ALTERNATIVE ESTIMATOR OF CLASSIFICATION EFFICIENCY; THE POINT DISTANCE INDEX (PDI)

In this section we propose the use of two alternative indices for use as figures of merit to be maximized as tests are sequentially selected for inclusion in a test battery. Both indices can be used to build a test battery by maximizing the efficiency of the LSEs used in the selection/classification process to achieve two different goals: maximizing PSE when only one composite is to be used; or maximizing PAE when using multiple composites for assignment to jobs.

The first index described is superior to $H_a$ for use in a selection process aimed at maximizing PSE and is referred to as "Max-PSE." The second index is called the point distance index, or PDI (Johnson,1970). We show that PDI is intuitively superior to $H_d$ for use in a test selection process directed at maximizing PAE. A rigorous proof of the superiority of PDI over $H_d$ most likely requires a model sampling experiment.

The Max-PSE index provides for maximizing the validity of the best single composite that can be obtained from any specified battery. "Best" is used in terms of the prediction of criterion scores in a combined sample that includes all the job samples. An operational test battery selected from an experimental test pool to maximize Max-PSE would necessarily provide a PSE of equal or higher value than could be provided by the use of $H_a$ in such a process.

The comparison of Max-PSE and $H_a$ is facilitated by the stipulation that means and standard deviations be equal for all variables across the job samples. The multiple correlation coefficients for the total sample (including all jobs) and for the first $k$ tests to be selected is, for each row of **F**, the square root of the summed squared values of the left most k columns of **F**. The average of these multiple correlation coefficients between the $k$ tests and the criterion variable corresponding to the $i^{th}$ row of **F** is, assuming our above stipulation holds, the validity of the best single predictor. The LSE that provides this value (Max-PSE) is the best composite for use in selection. The sum of squares of the same set of multiple correlation coefficients provides the value for $H_a$. It is the value of Max-PSE that should be used as the multiplier of the mean criterion score resulting from an optimal selection process, in order to provide the product that is equal to a MPP standard score, and thus provide a measure of PSE.

The formulae for Max-PSE and $H_a$ can each be written in terms of elements of **F**, where the elements **F** are defined as, $\mathbf{F} = (a_{ik})$, with $i$ identifying the LSE predicting the $i^{th}$ job criterion (corresponding to the $i^{th}$ row of **F**), and $k$ stands for the $k^{th}$ test to be selected (the $k^{th}$ column of **F** as built up in Horst's test selection procedure). Using this notation:

$$(\text{MAX-PSE})_k = 1/m \, \Sigma_i^m \left( \Sigma_j^k \, (a_{ij})^2 \right)^{1/2} \quad . \tag{3.4}$$

where $\Sigma_i^m$ is the summation over $i$ of the $m$ rows of **F**, and $\Sigma_j^k$, in this case, is the sum of a function of the elements of the $i^{th}$ row across $k$ columns of **F**. $(\text{MAX-PSE})_k$ is the figure of merit to be used in the test selection process for the selection of the $k^{th}$ test. Using this same notation $H_a$ is equal to $1/m \, \Sigma_j^m \left( \Sigma_j^k \, (a_{ij})^2 \right)$ .

In using the Max-PSE index as the figure of merit in a sequential test selection procedure, the first test to be selected will be the one with the largest average validity. Under our assumption of equal means and standard deviations across job samples, this test has the largest validity in the total sample and is clearly the one which should be used in the selection process, rather than the one with the largest squared validities as summed over the job samples. At this stage, the theoretical superiority of Max-PSE over $H_a$ is obvious. The second test to be selected is the one which provides, together with the first test selected, the largest average multiple correlation coefficient.

Since the rationale for the definition and proposed use of $H_d$ is based on psychometric rather than utility considerations, Horst made no claim as to the relationship of $H_d$ to a benefit measure such as MPP. A direct relationship does exist under the restrictive assumptions Brogden (1959) used for his model. However, there is no evidence that this or any similar relationship holds for a set of jobs and LSEs for which the validities are unequal across jobs. In the previous chapter we showed that $H_d$ values are more influenced by hierarchical classification effects than are MPP standard scores. This potential bias in $H_d$ could be controlled by the use of weights (i.e., $(\overline{R}\ /R_i)$) as described earlier.

However, there is another potential source of bias in $H_d$ for which such an intuitively helpful adjustment is not available. A difference in the evenness of the coverage of the joint predictor-criterion space affects $H_d$ and MPP differently. Thus the more uneven the coverage of this space, the less effective is $H_d$ as a predictor of MPP (i.e., PCE or PAE). We do not have the means of correcting $H_d$ for this latter type of bias but will propose an alternative index that will be more sensitive to the coverage of the joint predictor-criterion space.

Consider a hypothetical set of jobs for which half have coordinates clustered at two points in the opposite corners of the joint predictor-criterion space, and the other half are scattered over the remaining space relatively close to the midpoint. We will compare this first set with a second set of jobs that are scattered equally over all the regions of the joint predictor-criterion space, but each set retaining the same sum of squared distances from the midpoint. The two sets would thus both yield the same value for $H_d$ but provide quite different coverage of the space. Half the points in the first set lie on a single dimension, and it is these points that contribute the most to the value of $H_d$. It is intuitively attractive to believe that LSEs that are distributed more regularly over the joint predictor-criterion hyperspace would provide more PCE than if half of them were located on a single continuum stretching from corner to corner in that hyperspace. Jobs separated into two major families on the basis of their location on a single continuum permit hierarchical classification (but not allocation) effects, and $H_d$ is increased disproportionately as compared to MPP. If this intuitive logic is correct, it would be desirable to use an index that, compared to $H_d$, is more sensitive to evenness of coverage, or lack of coverage, with respect to the various regions in the joint predictor criterion space.

We presume that the average distance among job vectors, as measured in the joint predictor criterion space, is also a measure of how evenly this space is covered. The greater the average distance among job LSEs, the more even is this coverage for a given value of $H_d$, and the greater the potential for allocation efficiency. Also, it is highly desirable, as a matter of policy, to require every job to receive some benefit from the classification process; the index used as a figure of merit should, other things being equal, provide higher values when jobs are more evenly covered. The maximization of an average distance from the midpoint of the multidimensional space as contrasted with the sequential maximization of the squared distance measured on a single dimension (the dimension being considered for accretion), appears highly desirable.

A measure of average distance from the midpoint of the multidimensional space can be obtained by converting the intercorrelations among LSEs into measures of Euclidian distances and then using a multidimensional scaling procedure to obtain the Cartesian coordinates of the LSEs on the axis of the joint predictor criterion space. The maximization of the average interpoint distance in such a representation of the jobs, using a multidimensional scaling procedure, will provide the same numerical results as the use of PDI as described below.

The formula for PDI relates to $H_d$ much as the Max PSE index relates to $H_a$. Using the same notation as above:

$$\mathrm{PDI} = \Sigma_i^m \left( \Sigma_j^k (a_{ij} - a_j)^2 \right)^{1/2} \quad ; \tag{3.5a}$$

$$H_d = \Sigma_j^k \Sigma_i^m (a_{ij} - a_j)^2 \quad ; \tag{3.5b}$$

$$(H_d)_k = \Sigma_i^m (a_{ij} - a_j)^2 \tag{3.5c}$$

Note that all of formula 3.5a is used in making sequential selection decisions as to which test should be the next to be accreted, while only the rightmost sum in formula 3.5b, or formula 3.5c, is used to make this decision. $(H_d)_k$, as shown in formula 3.5c, is the index used to select the $k^{th}$ test to be added to the battery. In contrast, the comparable PDI index for making this decision, $(\mathrm{PDI})_k$, is simply PDI as shown in formula 3.5a. Thus the impact of using PDI instead of $H_d$ in the test selection process is best seen by comparing formulae 3.5a and 3.5c.

In using the PDI as the figure of merit for sequential test selection, the probability that tests will be selected other than those that would be selected by use of $H_d$ increases as the number of tests already selected increases. In general, PDI, as compared with $H_d$, will favor the accretion of tests that augment the differential validity of the LSEs with the smaller accumulated differential validities. These LSEs are those with smaller distances from the midpoint in terms of the already selected tests and appear to need a greater dimensionality to show a separation from the midpoint.

We intuitively feel that PDI is a better index for use in test selection than is $H_d$, but, of course, recognize that either a theoretical proof or empirical evidence is required before the substitution of PDI for $H_d$ can be recommended without reservations. We have initiated a model sampling periment to compare $H_d$ and PDI as predictors of both PAE and PCE, using a simulation approach that reflects real world data. We are planning a further model sampling experiment which will use hypothetical entities, predictors, and jobs designed to emphasize the differences between the two indices.

As with $H_d$, PDI can also be adjusted to eliminate hierarchical classification effects. The appropriate formula to eliminate these effects is as follows:

$$\text{Adjusted PDI} = \Sigma_i^m \left( \Sigma_j^k \left( \overline{R} / R_i \, (a_{ij}) - \bar{a}_j \right)^2 \right)^{1/2} \quad . \tag{3.6}$$

The rationale for this adjustment is the same as for the similar adjustment made to $H_d$. A weighting to reflect quotas can also be made in the same manner as for $H_d$.

PDI lacks the easy computational formula in terms of the matrix **C** and the convenient relationship to principal component (pc) type factor solutions that are provided by $H_d$. However, PDI has a direct relationship to multidimensional scaling; the axis produced in an initial multidimensional scaling solution can, like factor solutions, be rotated to more meaningful positions and can be used to identify job clusters and composites. This axis, as with the factors based on a maximization of $H_d$, can also be defined in terms of the predictor variables represented by $\mathbf{R_t}$.

In PDI we provide what we believe is an attractive alternative to the use of $H_d$ in test selection, an alternative aimed at the improvement of the resulting battery's PCE. PDI is proportional to PAE when the assumptions for Brogden's (1959) model are met. In contrast, the square of $H_d$ is proportional to PAE under the same conditions. It seems reasonable, although we have no definitive evidence as yet, that PDI is better related to PCE than is $H_d$, when the $R_i$ does not equal $\overline{R}$. As noted, the rationale for PDI is, as yet, currently intuitive, based on situational psychometric type evidence, rather than one based on utility. Still, we would tentatively recommend its use as an alternative to $H_d$; we expect soon to have model sampling results (in terms of MPP standard scores) to either support or refute this recommendation.

## D. TEST SELECTION STRATEGIES

The most effective battery for operational use for both selection and classification would include some tests selected by $H_d$ or PDI and some tests selected by $H_a$ or MAX-PSE. The presence of tests included to improve PSE will almost always increase the magnitude of the intercorrelations among job specific LSEs, but will not decrease the PCE of the battery and set of jobs for which these augmented LSEs are used.

Similarly, tests included to improve PCE cannot by their presence in the LSEs decrease PSE associated with the use of a single LSE selected to maximize predictive validity in the total job population. Neither will these tests that are best with respect to PCE decrease the PUE of a simultaneous selection-classification process, such as can be accomplished using the MDS algorithm. The inclusion of more tests will, of course, always raise the validities of the LSEs; more often than not, relatively low intercorrelations among the tests selected to improve PCE make these tests better than average prospects for improving PSE, although they are not necessarily the ones that would be selected in a sequential test selection procedure to maximize PSE.

It should not be necessary to include in a battery more than two or three tests selected to maximize $H_a$ or Max-PSE, nor more than seven or eight tests selected to maximize $H_d$ or PDI. If a smaller battery is to be administered to applicants for selection purposes and a larger classification battery administered to those who are accepted, the tests to be used for selection should first be removed from the experimental test pool and the tests for inclusion in the classification battery selected from the residuals; the classification tests should be selected using the residual relationships among the unselected experimental tests and criteria remaining after the effects of the tests selected to maximize PSE have been removed. Hopefully the test scores administered for selection can be added to those scores obtained for classification to form classification composites, since assignment of employees frequently involves accomplishing both selection and classification objectives.

If the test composites (e.g., aptitude areas) to be used as assignment variables are, by official policy and/or tradition, standardized so as to have the same mean and standard deviation for all composites, the test selection should reflect this intended usage. As described in a previous section, the row sums of squares that are to be aggregated to form $H_d$ or PDI should be adjusted using $(\overline{R}/R_i)$ as a multiplier. The use of this adjustment will hopefully prevent hierarchical classification effects from masking the PAE of tests being considered for selection. Even when the assignment process has been designed to capitalize on hierarchical classification efforts, as when the composites are LSEs with standard deviations proportional to $R_i$, it may be desirable to select at least a few tests using this adjustment. A model sampling experiment could determine the value of using this adjustment in the test selection process; the question of whether the closer relationship to PAE that is provided by this adjusted index will provide better utility when used in test selection requires further investigation.

Weighting the rows of **F** by the size of the quotas for the jobs corresponding to each row provides a means of emphasizing the comparisons that would be more numerous in the operational assignment process. A battery selected by a procedure that takes quotas into account should be used when the objective is to maximize the MPP standard score after an optimal assignment process has been accomplished. For a counseling situation where every comparison is considered to be equally important, it would be more appropriate to select tests without using weights that reflect job quotas.

Horst (1956b) illustrated a procedure for maximizing $H_d$ by assigning an optimal proportion of a fixed amount of testing time, and corresponding test length, to each test in an operational test battery. These proportions vary as the total battery testing time is changed. Within the time range used in three illustrations, the assigned times became increasingly different across tests, and the gain in differential validity increased, as the total battery time limit increased.

Horst (1956b) provides an iterative algorithm for successively improving the allocation of testing time (and test length) to increase the values of $H_d$. Horst's procedure requires the availability of data on testing time, reliability, intercorrelations and validities for all tests in a battery. Test length is assumed to have the same relationship to testing time throughout the range of testing times. Thus, given testing time, length, and reliability in one observed situation, test lengths and reliability are available for all other alternative time limits. Validities and intercorrelations of predictors for tests of any prescribed set of lengths are thus also functions of testing time and the validities and intercorrelations in the observed situation. Trial testing times that sum to the prescribed battery testing time will, in an iterative process, produce a value for $H_d$; the best set of testing times to maximize $H_d$ can be found by trial and error.

Horst's example in which he applied his algorithm used grade point averages for ten college subjects as the criteria and six cognitive aptitude tests as the predictor battery. The battery time limit for the observed situation was taken to be the sum of the time limits specified for the individual tests. In the first illustration the total time limits were halved. The total time limit was allowed to remain unchanged in the second illustration in which the total time (and length) was optimally allocated to the individual tests. The total time limit used in the second illustration was doubled for the third illustration.

Optimizing testing time increased $H_d$ by from 5 percent to 10 percent, with the larger gain accruing in the illustration with the largest total testing time. For these optimal testing times, the largest was ten times the size of the smallest, but none reduced to a time that approximated the effect of deletion.

Horst noted that no provision was made for test administration time. If administration time for each test had been added as a non-productive constant to the testing time required for the productive items, only the latter would have related to reliability, and thus to validities and intercorrelations among predictors. When the contribution of the item component for a shortened time limit could no longer compensate for the fixed administration time, test deletion would be indicated. Deletion would undoubtedly have occurred in Horst's example if he had included the effects of administration time.

A study was initiated to develop a computer program to simulate the building of a test battery from small increments of items (item blocks) from an experimental test pool (Johnson, 1970). Test selection from a battery represented by one block of items from each test was to be accomplished with the objective of sequentially maximizing $H_d$ at each step. What made this model different from standard sequential test selection procedures discussed earlier was that the first time a block was selected for accretion to the battery, a time charge for administrating the necessary directions was made against the allotted time. Thereafter an equivalent block could be selected as many times as it added more to $H_d$ than would the accretion of a block containing a new type of item that carried an administration time charge. The test selection process would halt when the desired total testing time, the sum of all administration and item times reached the desired value.

It was intended that selected batteries built block by block by the program would be checked against Horst's (1956b) algorithm modified to reflect administration time requirements. It was then intended that a model sampling experiment would be conducted to compare the effects on both PSE and PCE of using batteries selected to maximize $H_a$ and $H_d$ respectively. (Unfortunately, although the computer programming for this study was essentially accomplished, the study was not completed.)

The job sample used to conduct a test selection procedure is crucial to the development of a battery possessing high PCE. Jobs that span the joint predictor-criterion space of the population of jobs should be selected for use in this procedure rather than the jobs with larger quotas, or those deemed to have the greatest criticality. Job samples must be of adequate size to establish accurate estimates of validities, frequently making it desirable to under-represent large job families in order to over represent small job families.

The multidimensionality of the joint predictor-criterion space should be further enhanced by using several relevant criterion components for each job and the weighting of these components, as appropriate, accomplished differentially across jobs. The use of a single criterion component such as job knowledge or performance ratings will increase the probability that the criterion space across jobs is unidimensional, making it relatively difficult for PCE to exist, except for hierarchical *classification* effects that can be captured with a unidimensional predictor.

It is also essential to have an experimental test pool with heterogeneous content representing a number of factor domains such as: cognitive, traditional psychomotor abilities, video game skills, visual perception, performance under speed limits, and, especially, biographical, interest and self description measures. The cognitive domain should be represented by diverse content rather than by the relatively homogeneous measures of general mental ability found in the existing ASVAB. A preliminary screening of experimental tests to assure that only those with the highest predictive validity are included in the experimental pool can greatly reduce the effectiveness of test selection procedures intended to increase the PCE of the final battery.

Biographical, interest, and self description tests can be designed for differential prediction across jobs, or conversely, for the measurement of general adjustment, work related social skills, and motivation level. The latter generally predict supervisory ratings across all jobs, making such predictors better contributors to PSE than to PCE. Empirical keys for such tests are frequently highly correlated with general adjustment to the organizational environment, a measure that cuts across job families. This "g" factor in the non-cognitive domain is probably as prevalent as general mental ability is in cognitive tests. However, we believe it is easier to control "g" in the biographical, interest, and self description domain as compared with the cognitive domain. Johnson, Klieger, and Frankfeldt (1958a), and Johnson and Kotula (1958b) describe self-description tests designed to provide differential validity for a limited set of Army jobs by minimizing the "g" factor. Other techniques (e.g., forced choice items) to control an applicant's tendency to select the responses perceived to be socially desirable could also be used to control the non-cognitive "g" factor.

Between 1965 and 1975, information tests became very popular as a substitute for biographical and self-description tests. It was believed that such tests were more impervious to faking and more directly measured the positive consequences of interest and experience. Unfortunately, these tests tend to be indistinguishable from general mental ability in the joint predictor-criterion space. Thus, these "substitutes," while successful in certain instances where selection was the primary goal, have contributed considerably to the reduction of PCE in batteries, such as the ASVAB.

In summary, we believe the tools for selecting operational batteries with higher PCE from an experimental test pool should be used when more than one test composite is to be formed from the battery. However, we believe formal test selection from an experimental test pool must be preceded by carefully considered selection of measures for inclusion in such a pool. When this preliminary selection is based entirely on considerations of predictive validity, without thought of what might be needed to increase PCE, one should not expect significant gains in PCE, even when the further selection from the experimental pool maximizes PCE in the later test selection process. The formal test selection procedures cannot produce classification potential that was not placed in the experimental pool in the initial research step. Even with a wise selection of an experimental test pool, the test selection effort can be stalemated by the lack of an adequate criterion. The careful consideration of job criterion measures to avoid a unidimensional criterion space across jobs is also essential to a successful selection of a PCE rich battery.

## E. FACTOR ANALYTIC TECHNIQUES

We begin by considering how to use a weighted test composite which maximizes the value of $H_a$. Although a test composite which maximizes Max-PSE would have a theoretical superiority over one designed to maximize $H_a$, the difference is probably quite small. If policy specified the use of only one composite (one score per person to be classified), a close approximation to maximum performance is achieved by assigning the highest scoring persons to the job having the highest correlation with this composite, and so forth, just as with the single variable hierarchical classification model described in Chapter 1. A test composite with weights selected to maximize $H_a$ and used in this manner would be almost optimal. Thus a composite which corresponds to a factor in the joint predictor-criterion space which maximizes $H_a$, and is precisely defined as a weighted composite of the tests, closely approximates the characteristics desired in a single composite to be used in the same way AGCT was used by the Army.

The first principal component (pc) factor obtained in the joint predictor-criterion space will maximize factor contributions to $H_a$. We refer to this pc factor solution as $\mathbf{F}_a$. A pc factor solution of $\mathbf{C}$, or a derived pc solution obtained as an orthogonal rotation of $\mathbf{F}$, provides the same result. The latter is obtained by factoring $\mathbf{R}_t$ to obtain $\mathbf{F}_t$ and then extending $\mathbf{F}_t$ into the joint predictor-criterion space to obtain $\mathbf{F}$ a Dwyer factor extension solution, which in turn can be orthogonally rotated to a pc solution in the joint predictor-criterion space. Both methods sucessively maximize $\mathbf{H}_a$ as additional factors are added. In either case $\mathbf{C} = \mathbf{FF}' = \mathbf{F_a F_a}'$ and $V = \mathbf{FF_t}' = \mathbf{F_a F_t}'$. The pc solution derived from $\mathbf{F}$ has the conceptual advantage of being more directly linked to $\mathbf{R_t}$, making it easier to define each factor in terms of the tests.

$\mathbf{F_a}$ can be directly derived from $\mathbf{FA} = \mathbf{F}_a$, and $\mathbf{A}$ can be obtained by reducing $\mathbf{F'F}$ to a diagonal matrix of roots, $(D_a)$. A Grammian matrix such as $\mathbf{F'F}$ yields a unique solution for the matrix equation $\mathbf{A'(F'F)A} = \mathbf{D}_a$, where $\mathbf{A'A} = \mathbf{AA'} = \mathbf{I}$. Thus an algorithm for reducing $\mathbf{F'F}$ to a diagonal matrix yields precise values for $\mathbf{A}$ and $\mathbf{D}_a$. It is easily seen that $\mathbf{F}_a$, the principal component solution of $\mathbf{C}$ is equal to $\mathbf{AD}^{1/2}$. Also, since $\mathbf{F}_a$ is a pc solution, $\mathbf{F_a}'\mathbf{F}_a = \mathbf{D}_a$, where $\mathbf{D}_a$ is a diagonal matrix of successively maximized values of each factor's contribution to $H_a$ (they are, of course, also the eigen values or roots of both $\mathbf{F'F}$ and $\mathbf{C}$).

Factor scores for each individual pertaining to a factor for any orthogonal rotation of either $\mathbf{F_t}$ or $\mathbf{F}$ can be precisely defined as a sum of weighted test scores. An individual's factor score for the largest factor in the joint predictor-criterion space is defined as $\mathbf{Z}_1$. This score is equal to the $i^{\text{th}}$ person's row vector of test scores $(\mathbf{Y})_i$ multiplied by the corresponding column vector in a weighted matrix $(\mathbf{W})$. More generally, the complete factor score vector $(\mathbf{Z})_i$ is equal to $(\mathbf{Y})_i\mathbf{W}$, and $\mathbf{W} = (\mathbf{R}_t^{-1})\mathbf{F}_t\mathbf{A}$. To compute $\mathbf{W}$ without using an inverse of $\mathbf{R}_t$ (which may be very unstable for a large pool of tests), the following formula can be used: $\mathbf{W} = \mathbf{B}(\mathbf{D}_b)^{-1/2}\,\mathbf{A}$, where $\mathbf{D}_b$ is the diagonal matrix in the uniquely defined equation $\mathbf{B'R_tB} = \mathbf{D}_b$, and $\mathbf{B'B} = \mathbf{BB'} = \mathbf{I}$.

Another pc factor solution in the joint predictor-criterion space successively maximizes $H_d$. This factor solution, $H_d$, can also be derived as an orthogonal rotation of $\mathbf{F}$; $\mathbf{F}_a$ is equal to $\mathbf{FT}_o$ ; $\mathbf{T_o'T}_o = \mathbf{T}_o\mathbf{T}_o'$, and $\mathbf{T}_o'(\mathbf{F} - \mathbf{HF})'(\mathbf{F} - \mathbf{HF})\mathbf{T}_o$ is equal to a diagonal matrix of eigen values. As noted above, a unique solution for a matrix having the above properties, $T_o$, is readily available. A derivation and further explanation of $\mathbf{F}_a$, $\mathbf{F}_d$, and factor scores pertaining to both solutions, is provided in Appendix 3B.

The factor solution $\mathbf{F}_d$ has the same relationship to $H_d$ as $\mathbf{F}_a$ has to $H_a$. Just as the diagonal elements of $(\mathbf{F_a'F}_a)$ provide the successively maximized values of $H_a$ contributed by each factor, the diagonalized elements of $(\mathbf{F_d'F}_d)$ provide the successively maximized values of $H_d$ resulting from each factor.

Substituting $\mathbf{F}_d$ for $\mathbf{F}$ into the more general matrix formula for $H_d$, that is, $H_d = \mathrm{tr}((\mathbf{F} - \mathbf{HF})(\mathbf{F} - \mathbf{HF})')$, will yield the same value for $H_d$ using $\mathbf{F}_a$ as would be obtained using $F$, but in addition tr $(\mathbf{F_d'F}_d)$ and trace $(\mathbf{F_d}\ \mathbf{F_d'})$ are both equal to the total $H_d$ since $\mathbf{HF_d}$ is null (all column means of $\mathbf{F}_d$ are zero). While $\mathbf{F_dF_d'} = \mathbf{C}$, when no factors are dropped, if only a few factors are to be used (say, one to four that have the largest roots are to be retained), one can expect the approximation of $\mathbf{C}$ by $\mathbf{F}_d\mathbf{F}_d'$ to be relatively poor; in contrast, a very close approximation is provided by the first few factors of $\mathbf{F}_a$; $\mathbf{F}_a\mathbf{F}_a' = \mathbf{C}$. However, this better reproduction of $\mathbf{C}$ by $\mathbf{F}_a$, as compared to $\mathbf{F}_d$, is not relevant to classification efficiency.

The most compelling reason in this age of computers for using a few test composites, such as the nine aptitude area composites, instead of separate composites for each of the 30 to 40 job clusters recognized by the Army, or separate LSEs for the 260 Army jobs, is to provide understandability and creditability of assignment decisions to enlistees. Counselor recommendations and system decisions are frequently justified, or at least explained, in terms of test scores. Also the management system needs to record meaningful composite scores to determine an enlistee's eligibility on the basis of minimum standards required for requested job assignments. Such scores are also required to determine eligibility for various programs throughout his or her career.

Next we consider a hypothetical Army policy designed to meet the needs described above. This policy stipulates that counselors will be provided only four test composites rather than the undeniably more optimal thirty or so LSEs corresponding to currently existing major job clusters in the Army for use in accomplishing classification. These four composites are also intended to aid the counselor/classifier in providing career advice to the recruit. The first of our principal component factors of the type that will sequentially maximize $H_d$, the sums of squares of each column of $\mathbf{F}_d$ as defined above, would provide the best possible set of four composites for such a purpose.

The first (largest) factor of $\mathbf{F}_a$ could be implemented as a test composite and be used by itself as an assignment tool, as suggested earlier. Similarly, the use of several, say $k$, test components corresponding to the largest factors from $\mathbf{F}_d$, would provide the best classification efficiency obtainable from use of $k$ test composite scores. Such composite scores could be provided in profile form for counseling and as numbers for use in regression equations to predict performance on each job.

For example, to amplify this classification concept further, we will assume it has been decided to record only five test composite scores in the recruit's personnel file. These scores are to be used by the counselor/classifier in negotiating assignments with the recruit, and for later use in the determination of eligibility for various programs such as training courses and reenlistment. One of the five components should be equivalent to the largest (first) factor of $\mathbf{F}_a$. The other four composites should be selected to maximize classification efficiency. Since $H_d$ is the best known index we have for reflecting PCE, and the largest four factors from $\mathbf{F}_d$ maximize the magnitude of $H_d$ that can result from the use of any four factors (or composites), the four classification composites can reasonably be made equivalent to the four largest factors of $\mathbf{F}_d$. Each of these composites representing a factor is an LSE, with a factor for the dependent variable, the tests in the operational battery providing the independent variables.

To expand on our example of how five component scores could be utilized, we will describe an ideal situation for maximizing both creditability and PCE, and, consequently, utility. Assume that twelve tests are selected from a larger experimental test pool of 30 tests; 9 are selected to maximize $H_d$ , and three other tests are selected to maximize $H_a$. The intercorrelations of these 12 tests then become the $\mathbf{R}_t$ in the above development. $\mathbf{F}_t$ is computed and extended to the criterion space containing $m$ jobs to yield the Dwyer factor extension matrix, F. This $m$ by 12 matrix F is then orthogonally rotated to $\mathbf{F}_a$ with only the largest factor $\mathbf{F}_{a1}$ and the corresponding eigen vector, $\mathbf{A}_l$ being retained. Similarly the residual of F defined as $\mathbf{F_r}$, where $\mathbf{F_r}\ \mathbf{F}_r{}' = (\mathbf{C} - \mathbf{F}_{a1}\mathbf{F}_{a1}{}')$, is orthogonally rotated to produce $\mathbf{F}_{dr}$ and the largest four factors of $\mathbf{F}_{dr}$, and the corresponding four columns of $\mathbf{A}_{dr}$, retained for later use.

The four factors in the m by 4 matrix $\mathbf{F}_{dr}$ should be orthogonally rotated to more meaningful positions that correspond to simple structure with respect to the m jobs. Rotation to simple structure provides a structure across jobs and factors such that either high or low factor loadings are provided for most jobs in each job family. Using one of several available computer programs can accomplish this objective. Alternatively, a desired job structure could be reflected, as a hypothesis, in a surrogate **F** matrix, **L**, and used as a target matrix for the fitting of **FT** to the target matrix **L**. A formula for a transformation matrix, **T**, constrained to be orthogonal, that provides a least squares fit of **FT** to **L** is given by Green (1952). It may be desirable to adjust the rotated version of $\mathbf{F}_{dr}$ further, that is, $(\mathbf{F}_{dr})\,\mathbf{T}$, to form moderately correlated factors that provide a better fit to major job families.

A general factor score (for use in selection) and four differential factor scores (for use in classification) would be computed by using each individual's 1 by 12 row vector of 12 test scores, $(y)_i$, weighted by a **W** matrix to provide a 1 by 5 vector of factor scores, $(z)_i$, where $(y)_i\,W = (z)_i$. The least squares regression weights to be applied to the differential factor scores, to provide a best estimate (i.e., a LSE) of the criterion for the $j^{\text{th}}$ job, can be supplied for any orthogonal rotation of **F** as, $W_f = (\mathbf{R}_f)^{-1}\,(\mathbf{F}_{dr})\mathbf{T}$, where **T** is an orthogonal transformation matrix applied to obtain a more meaningful set of factors and $\mathbf{R}_f$ is the 5 by 5 matrix of intercorrelations among the 5 factors. ($\mathbf{R}_f$ will be the identity matrix if no oblique factor structure is introduced in the transformation of the axis to more meaningful positions). These regression weights could be converted to test composite profiles pertaining to each job. Profiles could be raised or lowered to reflect average job quotas (Cardinet, 1959).

A greater amount of PCE would result, in the above example, if all tests were used to compute the LSEs used inside the computer either to recommend or effect job assignments. However, it is highly probable that the LSEs based on a general factor score plus four differential factor scores would lose very little PCE as compared to the use of all 12 tests; a simulation study would be required to determine whether there would be a significant loss in classification efficiency.

The general factor score would be needed to reflect accurately profile level and as the basis for a minimum prerequisite (a cutting score) for entrance into highly technical school courses. This score would also be appropriate for use in selecting applicants for direct entry into the military for selected programs such as officer candidate school and helicopter piiot school.

There are other ways to make use of the $\mathbf{F}_a$ and $\mathbf{F}_d$ solutions described above. The selected example was provided to show that there are feasible ways to consider possible improvement of PCE in operational personnel systems, through the use of factor based, classification efficient, test composites.

## F. RESTRUCTURING JOB FAMILIES TO IMPROVE PCE

Job families are clusters of jobs in which each job can be presumed to be more similar to members of the same cluster than to members of the other clusters. The selection of a clustering procedure must consider both the measure of similarity and the process for determining number of groups, group membership, and, sometimes, membership criteria and group boundaries. If a means for addressing the latter two issues is provided, one has not just a clustering process and results but also a fully developed jobs taxonomy.

Most clustering algorithms either start with the most similar pairs and combine initial clusters and singlets into fewer and fewer clusters (leaf to stem) or start with the total group and successively separate clusters into more but smaller clusters as the process continues (stem to leaf). Multidimensional scaling and factor analysis provide a way of separating the total set of jobs into regions separated by hyperplanes. Multidimensional discriminant analysis provides another viable procedure for clustering jobs so as to assure they are more similar within than between categories.

Kruskal (1977) writes that the key difference between: (1) clustering algorithms that deal with similarity or proximity matrices and (2) multidimensional scaling, "is that multidimensional scaling provides a *spatial* representation for the proximities, while clustering provides a tree representation for them" (p. 29). Kruskal believes these two approaches are complementary, rather than competitive, with the latter more efficient when dissimilarities are small (as near group boundaries). Kruskal appears to be suggesting that boundaries could be more efficiently identified using multidimensional scaling and the fine tuning regarding the boundaries of families accomplished using a clustering algorithm. Numerous books have been written on clustering methodology [Hartigan (1975), Anderberg (1973), and Van Ryzin, ed. (1977)]. Numerical taxonomy is a related topic that is covered by another set of books including that of Sneath and Sokel (1973).

We are concerned with the clustering of jobs within the joint predictor-criterion space. Thus the measure of similarity or proximity that should be used in either a clustering algorithm or in alternative approaches (e.g., factor analysis), is the correlation among LSEs, if the goal of the clustering is to make selection (using LSEs) more efficient.

Alternatively, if the job clusters are intended to facilitate the classification process, the measure of proximity should be either the differences among the pairs of LSEs or a measure of the Cartesian distance among jobs represented as points in Euclidian space (e.g., our PDI).

An excellent example of using the correlations among LSEs as the measure of similarity for the clustering of jobs is provided by McLaughlin, et al.(1984). Their clustering was accomplished several ways on two independent cross-samples and the results compared. They concluded that the large cross-sample differences for their clustering results precluded their recommending the use of specific sets of clusters (job families) based on their empirical data. Because of the apparently low PCE of the ASVAB, it is doubtful that clustering on a measure of classification efficiency would have been more successful. However, clustering on $r_{ij}R_iR_j$ ($r_{ij}$ = correlation coefficient between $i^{th}$ and $j^{th}$ LSE, and $R_i$ = validity of $i^{th}$ LSE), might have produced a set of job clusters that would facilitate the effectiveness of hierarchical classification; clusters more homogeneous with respect to $r_{ij}R_iR_j$ would provide some increase in PCE (due to hierarchical classification effects), as well as improving PSE.

We believe the objective of clustering jobs into families for use with corresponding test composites for classification purposes should be to maximize either $H_d$ or PDI. We describe a procedure for maximizing $H_d$ but it would be easy to modify this approach to make use of PDI instead of $H_d$.

One approach to clustering would call for using the distance measures, the $p_{ij}$ values, as proximity measures and to select $p_{ij}$ sequentially, from smallest to largest of value for $p_{ij}$, and agglutinating[16] each pair of jobs that does not have a stronger connection to another job. The proximity of an agglutinated pair of jobs to other jobs or pairs could then be estimated as the average of the $p_{ij}$ that connects two of the evolving clusters. There are many varieties of this approach available for use, several have been implemented in off-the-shelf computer programs (some stem to leaf instead of leaf to stem). However, there is no reason to believe that those approaches would even approximate a maximization of $H_d$ in the completed set of clusters. In contrast to such approaches we describe a clustering

[16] The agglutination process is one of forming a new set which has a new meaning than that attached to either of the constituent sets; this new set has its own different relationships to other sets; the basic elements of the new set retains the separate identities of the basic set elements (of jobs) but the boundaries of the two constituent sets vanish in the agglutinated set. It is not accurate to describe agglutination as either a process of joining or linking, and two pairs of jobs are not merged. Thus the term "agglutination" was adopted.

algorithm that will sequentially maximize $H_d$ at each stage. Although relatively cumbersome, such an algorithm is entirely feasible in this computer age.

A preliminary step in our clustering algorithm is to create a matrix of squared differences among LSEs. This matrix, **D**, will have diagonal elements of zero and the remaining elements equal to the squared differential correlation coefficients between the $i^{th}$ and $j^{th}$ jobs (LSEs). This $m$ by $m$ matrix can be expressed as $\mathbf{D} = [d_{ij}]$. Horst's differential index, $H_d$, can be directly computed from this matrix since $H_d = (1'D1)/2m = (1/2m)\ \Sigma_i^m\ \Sigma_j^m\ d_{ij}$ . As clusters are formed these job families replace the individual jobs in their relationship to the rows and columns of **D**. Our clustering objective is to agglutinate jobs into families, reducing the order of **D**, while minimizing the reduction of $H_d$. To the extent that $H_d$ relates to PCE, a clustering procedure that maximizes $H_d$ for a prescribed number of clusters will also maximize the PCE for a given battery in a particular context of jobs and criteria.

The matrix of correlation coefficients, $\mathbf{R}_e$, among the LSEs and the factor extension matrix, **F**, is also required for the entire set of jobs on which the clustering process will be performed. The matrices **F** and **C** are as defined in this and the previous chapter. $\mathbf{R}_e$, the matrix of LSE intercorrelations, is equal to $(\mathbf{S}^{-1}\mathbf{C}\ \mathbf{S}^{-1})$, where $\mathbf{S}^2$ is a diagonal matrix comprised of the diagonal elements of the m by m matrix **C**, and $\mathbf{FF}' = \mathbf{C}$. The matrix **D**, discussed above, is computed from the elements of **F**, an $m$ by $n$ matrix. Letting $\mathbf{F} = [a_{ik}]$, and $\mathbf{D} = [d_{ij}]$, with $i$ and $j$ representing the row variables of **F** (jobs) as well as the row and column identifying an element of **D**, and $k$ the column variables (factors) of **F**, we can compute the elements of the $m$ by $m$ matrix **D**:

$$d_{ij} = \Sigma_k^n (a_{ik} - a_{jk})^2 \quad . \tag{9.6}$$

The smallest $d_{ij}$ will be selected at the beginning of each iteration. After the initial iteration, this selection will be made on a diminished **D** that has an order one less than the **D** of the previous iteration. At the end of each iteration $H_d$ equals ($\mathbf{1'D1}$). It is also necessary to adjust **F** and $\mathbf{R}_e$ during each iteration since an adjusted **F** is required to adjust **D**, and an adjusted $\mathbf{R}_e$ is required to adjust **F**. These three matrices as adjusted in the $g^{th}$ iteration will be referred to as $\mathbf{D}_g$, $\mathbf{F}_g$, and $\mathbf{R}_g$.

The average intercorrelation among the individual jobs in an evolving family will be stored in a column vector called $\mathbf{U}_g$. In each iteration the $p^{th}$ and $q^{th}$ elements of $\mathbf{U}_g$, $u_p$ and $u_q$, will be deleted and a new $s^{th}$ element added. The total number of jobs in each evolving job family, denoted as $n_i$ for the $i^{th}$ job or family, will also be stored. All $n_i$ have the value of one as the first iteration is commenced; all elements of $\mathbf{U}_g$ can also be appropriately initialized with values of unity.

$\mathbf{R}_{pq}$ is an $n_p$ by $n_q$ matrix consisting of all the cells of $\mathbf{R}_e$ that are correlation coefficients between elements (jobs) comprising the two job criterion variables agglutinated to form a new job family. Each coefficient, $r_{pq}$, is the correlation between the LSEs corresponding to the $p^{th}$ and $q^{th}$ job families. At first these coefficients are the same as those in $\mathbf{R}_e$, since initially all families consist of one job, but as jobs are agglutinated to form families of two or more jobs the new coefficients are computed using a correlation of sums algorithm. However, the elements of $\mathbf{R}_{pq}$ remain a selected set of the elements of $\mathbf{R}_e$.

The first iterative step of the algorithm is to select the smallest numerical value of $d_{ij}$ and to agglutinate the two corresponding jobs, or job families. At the start of the first iteration the rows and columns of $\mathbf{D}$ will all correspond to jobs, but in later iterations one or both rows and columns corresponding to a $d_{ij}$ may represent evolving job families. When the smallest $d_{ij}$ is identified, the row (i.e., the specific value of $i$) is designated as $p$ and the column (i.e., the specific value of $j$) is designated as $q$. If the $q^{th}$ job family contains $n_p$ jobs and the $q^{th}$ job family contains $n_q$ jobs, $\mathbf{R}_{pq}$, as described above, is an $n_p$ by $n_q$ matrix and there is a product moment coefficient, $r_{pq}$, corresponding to the selected $d_{ij}$.

The $(n_p + n_q)$ by $n_j$ matrix, also consisting of cells from $\mathbf{R}_e$, is denoted as $\mathbf{R}_j$. This matrix, $\mathbf{R}_j$, consists of the correlation coefficients between each member of the set of $p + q$ job criteria and each member of the $j^{th}$ job family. There will be a separate $\mathbf{R}_j$ for each of the $(m - g)$ criterion variables remaining after the two criterion variables associated with the $d_{ij}$ selected in step one of each iteration.

The following steps in each iteration provide for the elimination of the two rows in $\mathbf{F}_g$ and the row and column in $\mathbf{R}_g$ and $\mathbf{D}_g$ associated with the last selected $d_{ij}$. This is followed by the computation of a new row of $\mathbf{D}_g$, $\mathbf{F}_g$, and $\mathbf{R}_g$ and the corresponding column for $\mathbf{D}_g$ and $\mathbf{R}_g$. Only one element of $\mathbf{U}_g$, one row of $\mathbf{F}_g$, and one row and corresponding column of $\mathbf{D}_g$ and $\mathbf{R}_g$ is recomputed during each iteration.

The iterative steps for this algorithm are as follows:

(1) Select smallest non-diagonal $d_{ij}$ and identify the corresponding $p^{th}$ and $q^{th}$ rows of $\mathbf{F}_g$ and both the $p^{th}$ row and $q^{th}$ column of $\mathbf{D}_g$ and $\mathbf{R}_g$.

(2) Compute a new $s^{th}$ row and column of $\mathbf{R}_g$ to replace the $p^{th}$ row and $q^{th}$ column, both of which will be deleted from $\mathbf{R}_g$; $s$ is equal to $m - g$. This new

$s^{th}$ row consists of correlation of sums coefficients between the sum of all job criterion scores comprising the $p^{th}$ and $q^{th}$ job families, and each of the remaining variables corresponding to the rows of $\mathbf{R}_g$ (i.e., all variables except $p$ and $q$). $r_{sj} = r_{(p+g)j} = (\mathbf{1'R}_j\ \mathbf{1})/((L_p{}^2 + L_q{}^2 + r_{pq}\ L_p\ L_q)^{1/2}\ L_j)$; $L_i = (n_i + ni(n_i - 1)u_i)^{1/2}$. $\mathbf{R}_{g+1}$ is created by deleting the $p^{th}$ row and $q^{th}$ column of $\mathbf{R}_g$ and then bordering this matrix with the row vector $(r_{sj})$ and the column vector $(r_{sj}) = (r_{is})'$.

(3) Compute a new $s^{th}$ row of $\mathbf{F}_g$ to replace the $p^{th}$ and $q^{th}$ rows of $\mathbf{F}_g$ and border $\mathbf{F}_g$ with the row vector $(a_{sj})$; the $j^{th}$ element of the $s^{th}$ row vector $(a_{sj})$ equals

$$(a_{pj}\ L_p + a_{qj}\ L_g)/(L_p{}^2 + L_q{}^2 + r_{pq}\ L_p\ L_q)^{1/2} \quad .$$

The term $r_{pq}$ is equal to $(\mathbf{1'R}_{pq}\ \mathbf{1})/(L_p\ L_q)$. This new matrix, $\mathbf{F}_g$ with the $p^{th}$ and $q^{th}$ rows deleted and $\mathbf{F}_g$ then bordered by the vector $(a_{sj})$, is denoted as $\mathbf{F}_{g+1}$.

(4) Delete $u_p$ and $u_q$, and add a $u_s$ element to the vector $\mathbf{U}_g$.

$$u_s = [L_p{}^2 + L_q{}^2 - n_p - n_q + 2\ (\mathbf{1'R}_{pq}\ \mathbf{1})]/[(n_p + n_q)^2 - n_p - n_q] \quad .$$

(5) Compute new $s^{th}$ row and column for $\mathbf{D}_g$; Using $\mathbf{F}_{g+1}$, $d_{sj} = \sum_i{}^s(a_{sj} - a_{ij})^2$. Delete $p^{th}$ row and $q^{th}$ column of $\mathbf{D}_g$ and border this resulting matrix with the row vector $(d_{sj})$ and the column vector $(d_{is}) = (d_{sj})'$.

(6) Compute $H_d = (\mathbf{1'}\ \mathbf{D}_g\ \mathbf{1})/2(m - g)$, for the $g^{th}$ iteration and compare with the values of this index obtained in step 6 of the previous iteration; consider the number of job clusters ($m$), and trend in values of $H_d$ to decide whether to stop or to start another iteration (steps 1 through 7).

(7) Prepare to commence the next iteration by adding one to $g$. This updating is accomplished as follows: (a) the $\mathbf{R}_{g+1}$ computed in step 2 is now $\mathbf{R}_g$; (b) the $\mathbf{F}_{g+1}$ computed in step 3 is now $\mathbf{F}_g$; (c) the $\mathbf{U}_{g+1}$ computed in step 4 is now $\mathbf{U}_g$; (d) the $\mathbf{D}_{g+1}$ computed in step 5 is now $\mathbf{D}_g$.

At the conclusion of the clustering process most analysts will wish to recompute $\mathbf{V}$ and $\mathbf{C}$. The $\mathbf{F}$ matrix for the final set of job families is the last adjusted $\mathbf{F}_g$; $\mathbf{F}_g\ \mathbf{F_t}' = \mathbf{V}_g$, and $\mathbf{F}_g\ \mathbf{F_g}' = \mathbf{C}_g$. If the empirically determined job families are to be used operationally, a test selection process in which $\mathbf{V}_g$ is substituted for $\mathbf{V}$ should be accomplished.

If the jobs (LSEs) were graphically plotted on two-dimensional projections from the joint predictor-criterion space, we would expect half or more of these points to be as close to the hyperplanes separating families as to the centroid of the family. We have no reason to believe that jobs will, in general, cluster in this space more densely near the centroids than near the boundaries of traditional job families. We can, of course, capitalize on chance and locate our separating hyperplanes through less dense regions, but must expect most of the benefits of such fitting to disappear in independent cross-samples.

Any structure devised to cluster over one hundred jobs into less than a dozen families will necessarily include jobs in a family that are much more similar to certain jobs in other families than they are to the more representative jobs in their own job family. Only the core jobs can be expected to yield good results when classification reliability is assessed using independent cross-samples. Thus, one must not expect a great deal of reliability for clustering results in cross-sample comparisons unless jobs close to boundaries are weighted less, and/or misclassifications of such jobs to job families with proximal boundaries are weighted less than disagreements as to the classification of the core jobs of each family.

Both the dimensionality of the joint predictor-criterion space and the relationships among jobs in this space can be explored using the factor approaches discussed in the last section. If it is desired to view the relationships among jobs in the smallest possible number of dimensions, the pc solution of **C** (referred to as $\mathbf{F}_a$ in the previous section) should be used. If a solution in terms of factors which have relevance to PCE is desired, $\mathbf{F}_d$ should be used. The rotation of $\mathbf{F}_d$ to simple structure would aid in the identification of major job families that can be appropriately utilized in the classification process; each rotated factor can be defined as a test composite (a LSE, based on all tests in the battery, in which the dependent variable is the rotated factor).

The rotation of a pc type factor solution to aid in the classification of n jobs into families can be accomplished by using a target factor structure that represents either a hypothesis or the results of a clustering procedure. As described in the previous section, the pc type solution, itself an orthogonal rotation from the factor extension matrix, **F**, can be orthogonally rotated to provide a least squares fit to the target matrix (Green, 1952). Boundaries between job families can then be located graphically and other data considered in the classifying of jobs located near these boundaries.

One appropriate hypothesis for reflection in a target matrix could be obtained through the use of a clustering approaach to identify the core jobs of major families. The least squares fit to a target matrix could then be accomplished using only these core jobs to define the target matrix. The orthogonal transformation matrix obtained from accomplishing this fit could then be applied to the remaining rows of $\mathbf{F}_d$ and the graphical consideration of family boundaries accomplished. However, this orthogonal rotation of $\mathbf{F}_d$ could be first "fine-tuned" by hand rotations to improve simple structure. Other hypotheses could be formed and implemented in a target matrix from consideration of existing officially imposed clusters of jobs or the structure implied by the location of job relevant school courses in the same school or school department.

There is no real problem in having a target factor structure that exceeds the dimensionality of the joint predictor-criterion space. For example, the points actually located on a plane can be assigned coordinates in a three-dimensional (or higher) space by tilting the plane so at least some of the points will have non-zero coordinate values on all axes. This tilting of a space within a larger space permits a greater flexibility in locating axes through swarms at points to increase the quality of simple structure. Since each of these axes can be defined as a test composite and used in a personnel assignment process, the use of the additional axes can improve PCE.

Although it has been proposed by Sokel (1977) that factor analysis and/or multidimensional scaling methods be used to identify major proximity differences, and clustering be used to measure smaller differences, we propose using clustering to help form hypotheses and the spatial methodologies to locate boundaries. The latter methodology is more amenable to the consideration of other data and policy constraints than the more numerical clustering approach. Also, it is desirable to have the final job classification process result in a factor solution, since these factors can be precisely duplicated by a regression equation of predictors usable as the test composites in the assignment process.

The increase in PCE that can be obtained by increasing the number of job families must not be confused with the number of jobs in Brogden's model (1959). In the latter, each additional job is assumed to be accompanied by an additional dimension in the joint predictor-criterion space. The improvement of PCE through the increasing of the number of job families does not depend on this assumption. Adding other jobs that are distributed throughout the same space, with the same density, and same average distance from the midpoint as the existing jobs, while retaining the same number of families, will not improve PCE. In contrast, increasing the number of job families and corresponding composites increases PCE all the way to the maximum number, where the number of jobs equals the number of families.

## G. SUMMARY AND CONCLUSIONS

The primary object of this chapter is to show how to depart from the ideal process for realizing classification efficiency while minimizing loss of PCE. The ideal process requires the use of a separate LSE as the assignment variable for each job, and the use of all predictors (i.e., with no test selection) to compute each LSE; there is no job clustering and no use of test composites other than LSEs in the ideal process for maximizing PCE. Thus selecting a subset of predictors for use in a classification test battery, the reduction in the number of jobs or job families (a result of clustering), and the definition of test composites other than LSEs for use as assignment variables all represent departures from the ideal.

The addition or deletion of tests, or jobs in a job family, as one step in the development of a classification system, should be based on the maximization of a figure of merit directly related to PCE; PCE will not be maximized by test selection or job clustering that seeks to maximize predictive validity. To improve PCE, decisions concerning the content of the: (1) experimental test pool, (2) operational battery, or (3) test composites used in the assignment process, must be made with improvement of PCE specifically in mind; PCE can be expected to be reduced as a consequence of actions taken to improve PSE when these actions are departures from the ideal process that is optimal for both selection and classification.

Departures from the ideal selection and classification process may be required to keep testing time within practical bounds and to provide a practical number of test composites, and corresponding job families for use by recruiters and counselors in the initial acquisition and assignment process. A smaller number of composites (with matching job families), as compared to the ideal number, may also be required by administrative restrictions or from lack of adequate validity data that together prevent the use of the ideal process. There are also other requirements for a smaller number of test composites related to relatively homogeneous job families, including the need for such convenient predictive scores in establishing minimum prerequisites for entry into programs occurring later in a soldier's (or worker's) career. Thus techniques for maximizing PCE in test selection, test composite identification, and clustering jobs into families are valuable tools, albeit they are describing a "best" way for departing from the ideal process.

Horst developed test selection procedures that consider criterion measures for multiple jobs: two maximize absolute validity, $H_a$, to improve the PSE of the selected battery (Horst, 1955, 1956b); two others maximize differential validity, $H_d$. Of the latter, one uses an accretion[17] algorithm (Horst, 1966) and a second uses a deletion algorithm (Horst, 1960).

---

[17] The sequential addition of tests to a battery is consistently referred to in the psychometric literature as the "accretion process."

A much greater potential contribution to utility than obtainable from test selection alone can be provided by selection of an optimal administration time, and, by inference, test length or number of item, for each test in a battery or experimental test pool. This technique permits the tailoring of tests to provide a near maximum classification efficiency within the total time limits allotted to the administration of an operational battery. To this end Horst provided an algorithm to maximize $H_d$ (Horst, 1966) and another to maximize $H_a$ (Horst, 1956c). We provide an approach for adding the consideration of administration time, permitting the reconstitution of tests to form a battery, while maximizing $H_d$ in accordance with Horst's algorithm.

We suggest the point distance index (PDI) as an alternate figure of merit to be maximized in selecting tests because we believe there is a closer relationship between PDI and PCE than between $H_d$ and PCE under the most commonly occurring conditions (i.e., when the assumptions of Brogden's (1959) model have not been met). $H_d$ is proportional to the square of PAE, and PDI is proportional to PAE itself when these assumptions do hold.

The clustering of jobs into families is to take a big step away from the ideal process, primarily because each test composite (hopefully an LSE) used as the assignment variable for all jobs in a family does not approach the accuracy with which each job in the family is represented by its own LSE. However, the personnel system requires test composites for: (1) counseling, (2) setting visible minimum prerequisites for training courses, and (3) both controlling reassignments at later career decision points and providing job incumbents with career relevant information. Rather than focusing on job clustering, one should concentrate on the matching of jobs to test composites so as to maximize $H_d$, $H_a$ or alternative indices. To this end we describe factor solutions that maximize $H_a$ for any given number of factors and another solution which similarly maximizes $H_d$. These factors can be rotated to provide a match between factors and jobs, and then precisely defined in terms of the predictor tests.

The value of the methods suggested for obtaining (unfortunately, the verb gleaning is frequently more descriptive of what is required) the available PCE from an experimental test pool, in the context of a special set of jobs and criterion measures, depends on the skill of the researcher in developing predictor and criterion variables to be used in creating the experimental data. The validity generalization movement has provided a great service in pointing out the difficulty of obtaining PCE. However, it is inappropriate to suggest that the joint predictor-criterion space is inherently unidimensional in nature until a concerted, technically correct, effort is expended[18] with the goal of maximizing PCE in both the development and selection of measures for inclusion in the experimental pool. Batteries developed to maximize PSE and validated against limited unidimensional job criteria are not the appropriate reference points concerning the feasibility of an effective allocation process. We believe that there is a strong potential for the identification of several additional dimensions in the joint predictor-criterion space whose existence can be confirmed with the concern and care used by Hunter (1986) with respect to the existence of general mental ability, clerical speed, and psychomotor ability in the joint GATB-criterion space.

The predictor space should never be substituted for the joint predictor-criterion space in the determination of composites or job families to be used for classification. Equally important, an index closely related to PCE, rather than to PSE, should be used to make these determinations.

In this chapter, approximately half of the recommended methodologies for increasing PCE were developed by Horst. The remainder, including Max-PSE, PDI, the particular applications of factor analytic approaches, and job clustering to maximize $H_d$, appear here in this chapter for the first time. We hope that with more techniques and with the linking of Horst's and Brogden's contributions, more investigators will make a deliberate effort to improve the PCE of a battery that is to be used to accomplish classification.

The maximum PCE for a battery is obtained when separate LSEs, each based on the full number of available tests (e.g., the experimental test pool), are provided for each job. The reduction of tests for an operational battery, the use of a smaller set of composites, or the merging of jobs into families all represent departures from the ideal. These departures should be made so as to minimize the loss of PCE as compared to the ideal process. This can be accomplished by selecting tests, and either using a separate LSE for each job or selecting composites and jobs for inclusion in families, using procedures that consider the effect on PCE as tests are selected, composites formed, and jobs or evolving job clusters are merged into job families used in the classification process.

---

18 Most would agree that one should always make a heroic effort to find a difference before accepting the null hypotheses and a super-heroic effort before concluding that the null hypothesis has been proven. Concluding that there is only one relevant dimension, general mental ability, is at least equivalent to accepting a null hypothesis, and in the eyes of some, equivalent to concluding that the null hypothesis has been proven.

The accomplishment of accretion or deletion of tests or jobs should be based on a figure of merit directly related to PCE; PCE will not be optimized by test selection or job clustering that seeks to maximize predictive validity. To improve PCE, one must make decisions in the test/battery/composite development process designed to improve PCE rather than aiming at an improvement of PSE and hoping that PCE will be improved as a side effect.

Horst has provided a number of test selection procedures that simultaneously consider multiple criteria (one for each job). Two of these maximize absolute validity, $H_a$, and are most useful in the improvement of the PSE of a battery (1955, 1956c). One provides an accretion process (1966), and another a deletion process (1960) for test selection to maximize differential validity ($H_d$). The time and length for each of a set of tests already selected for inclusion in a battery is considered in one algorithm to maximize $H_d$ (1966), and another algorithm (1956c) to maximize $H_a$. We have provided the point distance index (PDI) as an alternative figure of merit to be maximized in selecting tests or designation of testing times. We believe PDI is more closely related to PCE than is $H_d$ when the assumptions of Brogden's (1959) model have not been met.

The clustering of jobs into families may require a decision as to the relative priorities of maximizing PSE and PCE. A different structure could result from the agglutinating of jobs with high correlations among LSEs (to maximize PSE) as compared to the agglutinating of jobs with small differences between LSEs (to maximize PCE). All competition between the two objectives disappear as families become so small that every job is represented by its own LSE.

While we can eliminate the need for clustering jobs into families in the assignment process by using the full regression equations as the assignment variables, the personnel process will still require test composites for counseling, the setting of visible minimum prerequisites for school courses, and controlling entry into MOSs and special programs. The effective use of these test composites may require the identification of job families for which one or more of the test composites have special relevance. We have described factor solutions, one which maximizes $H_a$ for a given number of factors and another which similarly maximizes $H_d$. These solutions can be rotated to provide factors with meaningful relationships to jobs and/or job families, and then completely defined in terms of the predictor tests.

We recognize that ultimately the value of the methods provided above for gleaning the available PCE from an experimental test pool, in the context of a special set of jobs and criterion measures, depends on the sk'll of the research psychologist in the creating of predictor and criterion variables. The validity generalization movement has provided a great service in pointing out the difficulty of this task. However, no one can legitimately say that the joint predictor-criterion space is inherently unidimensional in nature until heroic efforts have been made, both in the development of measures and in paying deliberate attention to PCE in making decisions about batteries, composites, job families, and the selection/asssignment process. It would not be good science to examine batteries developed to maximize PSE, and validated against limited unidimensional job criteria, to reach conclusions concerning the feasibility of classification.

It is also not true to the scientific method to attend only to the predictor space in the determination of composites to be used for classification. Nor does it aid the classification process to cluster jobs (or treatment categories) in any domain other than the joint predictor-criterion space. Hunter (1986), using data that relies primarily on the ratings of supervisors, has concluded that the GATB contributes to three dimensions in this space. We believe that there is potential for more than these three, but their existence and usefulness should, at least eventually, be established with the concern and carefulness used by Hunter to confirm the existence of general mental ability, clerical speed, and psychomotor ability in the joint GATB/criterion space. We believe the methodology provided in this publication, including the model sampling techniques described in the following chapter, should be helpful to the research that needs to be accomplished on both the potential and the existing operational utility of classification.

# APPENDIX 3A
# ELIMINATING TESTS FROM COMPOSITES, AND/OR TESTS FROM BATTERIES, WHILE MINIMIZING LOSS OF CE

Brogden (1964) described an approach for the elimination of predictor variables whose additional contribution to classification efficiency (CE), beyond that provided by the retained variables, is zero or negligible. He was concerned with the elimination of variables from the FLS composites associated with each job family, rather than in the selection of tests for inclusion in an operational battery. However, the two concepts are similar in that tests eliminated for all composites would also be thus identified as not needed in the battery.

In the approach described by Brogden, regression weights for FLS composites make up one row of the matrix **W**; the columns of **W** represent predictor variables and the rows correspond to jobs. Brogden pointed out that classification efficiency is unaffected by the addition or subtraction of constants to a column of **W**. The addition of a constant which reduces all the weights for a predictor, i.e., for one column, to zero has the effect of eliminating that variable from all composites (and thus from the battery).

We would not expect all elements in a column of **W** to be reduced to zero in an analysis of empirical data. Some degree of closeness to zero would be established as either equivalent to zero or too small to make more than a trivial contribution. Closeness to zero could be measured using various metrics and criteria for making the decision. The average absolute distance from the column mean, the standard deviation, or the range of column values could be proposed as candidate metrics.

While Brogden did not propose a metric to be used in measuring how close to zero columns of **W** can be reduced, and he certainly did not suggest that the columns of **W** could be rank ordered with respect to their closeness to zero after the optimal selection of column constants, his basic concept can be related to Horst's and MacEwan's elimination method of selecting tests (1960). We can see this similarity by noting that **W** can be depicted as a triangular factorization of $\mathbf{R}_t$ (i.e., $\mathbf{F}_t$) extended to **V** to obtain $\mathbf{F}_V$, in our notation, and this $\mathbf{F_v}$ equated to $\mathbf{W}$. If each predictor variable is depicted, in turn, as the last column of $\mathbf{F_v}$ and the standard deviations of each variable while in last place compared, the elimination of the variable with the smallest standard deviation would provide the same result as using the algorithm proposed by Horst.

Thus we see that if we start with Brogden's concept and apply Horst's metric (the squared standard deviation of the elements of a column), and adopt Horst's concept of looking for the variable making the smallest contribution (contrasted with Brogden's search for a variable making *no* contribution), we have conceptually arrived at Horst's elimination method. It is easy to see that identifying the variable for which the standard deviation of the regression weights applied to the component of a variable orthogonal to all other variables will also identify the variable whose elimination will minimize the reduction of $H_d$.

The algebraically equivalent solution to that obtained by computing $m$ separate triangular $F_v$ solutions, each solution placing a different test variable in last place, is more economically obtained by using Horst's formula:

$H_d = \text{tr } \mathbf{C_p} - (\mathbf{1'}\ \mathbf{C_p}\ \mathbf{1})/m$. The equivalent of identifying the variable with the smallest regression weights after minimizing these weights by subtracting the appropriate constant (i.e., the mean value), is obtained by retaining the variables defining $\mathbf{C_p}$ which provide the largest values of $H_d$ *defined a*s a function of $\mathbf{C_p}$.

The selection of tests for inclusion in test composites smaller than FLS composites requires a different strategy for the selection of tests than is appropriate for the selection of an operational battery. Tests removed from one composite can remain in other composites. Thus, Brogden's objective in his 1964 article relates to the classification efficiency of operational composites as contrasted with the potential classification efficiency of an operational battery, the goal of Horst's DV approach.

Just as Brogden (1964) did not directly offer a means for selecting an operational battery from an experimental test pool, Horst did not publish a method for eliminating the less productive, least classification efficient, tests from some FLS composites but not from others. We suggest that the identification of tests which could be appropriately left out of FLS composites could be accomplished as a byproduct of an accretion method for selecting tests to maximize $H_d$. As each successive test is selected, any job whose validity is equal to the mean value of the just computed column of $\mathbf{F_v}$ can have that test (the test corresponding to the just computed column) eliminated from the composite for that job without appreciably reducing the value for $H_d$ (or $H_p$).

# APPENDIX 3B
# A FACTOR SOLUTION FOR MAXIMIZING $H_d$

In this appendix we provide a development of two factor solutions in joint predictor-criterion space comparable to a PC solution of $R_t$, except that $H_a$ and $H_d$ are respectively maximized instead of the maximization of factor contributions in test space. A PC solution of $\mathbf{C}_p$ provides factors for which the factor contributions are successively maximized considering only the criterion (job) variables as the dependent variables and the predicted performance, in terms of the test variables, as the independent variables.

For a PC solution of $\mathbf{C}_p$, the covariances among predicted performance estimates, the following relationships hold: $\mathbf{F}_c\mathbf{F}_c' = \mathbf{C}_p$, $\mathbf{F}_c'\mathbf{F}_c = \mathbf{D}_c$, where $\mathbf{D}_c$ is a diagonal matrix of eigen values from the equation $\mathbf{A}_c'\mathbf{C}_p\mathbf{A}_c = \mathbf{D}_c$, and both $\mathbf{A}_c\mathbf{A}_c'$ and $\mathbf{A}_c'\mathbf{A}_c$ equal the identity matrix. The sum of the diagonal elements of $\mathbf{D}_c$ are equal to Horst's absolute validity index, $H_a$. Conventionally, these eigen values which equal the contribution of each factor to $H_a$ are listed in order of magnitude, from left to right, and the contribution of $k$ factors to $H_a$ is maximized for $k$ factors by selecting the first $k$ factors on the left.

An investigator can maximize $H_a$ for $k$ factors by directly computing $\mathbf{F}_c$ as $\mathbf{A}_c\mathbf{D}_c^{1/2}$, and then selecting the $k$ factors with the largest factor contributions, or by converting $\mathbf{F}_v$ to $\mathbf{F}_p$ ($\mathbf{F}_p = \mathbf{F}_c$) using $\mathbf{F}_p = \mathbf{F}_v\mathbf{A}_p$ , where $\mathbf{A}_p'(\mathbf{F}_v'\mathbf{F}_v)\mathbf{A}_p = \mathbf{D}_p$, and selecting the $k$ factors having the largest factor contributions. Before pruning to $k$ factors, $\mathbf{F}_c$ will have $m$ columns (i.e., factors) while $\mathbf{F}_v$ will have as many factors as there are tests (i.e., $n$). However, the non-zero columns of $\mathbf{F}_p$ will equal $\mathbf{F}_c$ regardless of whether $n > m$, $m > n$, or $n = m$.

As discussed in Appendix 2B, various transformation matrices $\mathbf{T}$ can be used to transform both $\mathbf{R}_t$ and $\mathbf{V}$ to factor solutions that provide the same factors for both $\mathbf{F}_t$ and $\mathbf{F}_v$. That is,

$$\begin{bmatrix} \mathbf{R}_t \\ \cdots \\ \mathbf{V} \end{bmatrix} \mathbf{T}_1 = \begin{bmatrix} \mathbf{F}_t \\ \cdots \\ \mathbf{F}_v \end{bmatrix} .$$

We can commence with $\mathbf{F_c}$ and extend this solution to $\mathbf{F_t}$. This factor extension process can be expressed as

$$\begin{bmatrix} \mathbf{R} \\ \cdots \\ \mathbf{V} \end{bmatrix} \mathbf{T_2} = \begin{bmatrix} \mathbf{F_{tc}} \\ \cdots \\ \mathbf{F_c} \end{bmatrix},$$

with $\mathbf{T_2} = \mathbf{T_1 A_p} = \mathbf{A_t D_t}^{-1/2}\, \mathbf{A_p}$, and using $\mathbf{F_v} = \mathbf{V A_t D_t}^{-1/2}$,

$$\mathbf{A_d}'(\mathbf{F_v}'\mathbf{F_v})\mathbf{A_d} = \mathbf{D_c}.$$

The investigator can choose other solutions for $\mathbf{F_v}$, thus changing the above formula for $\mathbf{T_2}$. We chose to use the expression: $\mathbf{T_1} = \mathbf{A_t D_t}^{-1/2}$. Factor scores corresponding to the factors represented by the columns of $F_c$ constitute the elements of $\mathbf{Q_c}$, a matrix that can be computed by the formula $\mathbf{Q_c} = \mathbf{Y}\, R_t^{-1} F_{tc}$.

The equality of $\mathbf{F_c} = \mathbf{F_v A_p}$, as used above, follows from the well known theorem that for a positive semi-definite matrix $\mathbf{M}$, the equation $\mathbf{A_m}'\mathbf{M}\, \mathbf{A_m} = \mathbf{D_m}$ is uniquely defined in that there is only one orthogonal (or orthonormal) matrix, $\mathbf{A_m}$, and only one diagonal matrix, $\mathbf{D_m}$, that can fulfill this equation. Further, a factor matrix, $\mathbf{F}$, defined as $\mathbf{F} = \mathbf{F_v A_p}$ must be a PC solution if $\mathbf{F'F}$ equals a diagonal matrix, $\mathbf{FF'} = \mathbf{C_p}$, and $\mathbf{C_p}$ is positive semi-definite. It is evident that $\mathbf{F_v A_p}$ is a PC solution of $\mathbf{C_p}$ and must be equal to $\mathbf{F_c}$. Thus, an investigator has the choice of directly factoring $\mathbf{C_p}$ to obtain $\mathbf{F_c}$ ($\mathbf{F_c} = \mathbf{A_c D_c}^{1/2}$ where $\mathbf{C_p} = \mathbf{A_c D_c A_c}'$), or can extend $\mathbf{F_t}$ to $\mathbf{V}$ and obtain $\mathbf{F_v}$, a factor solution that can be transformed to $\mathbf{F_p}$, using the relationship, $\mathbf{F_p} = \mathbf{F_v A_p}$; $\mathbf{F_p} = \mathbf{F_c}$.

Horst's differential validity index, $H_d$, has the same relationship to the matrix , $\mathbf{G} = (\mathbf{F_v} - \mathbf{HF_v})$, as $H_a$ has to $\mathbf{F_v}$. We note that $\mathrm{tr}\,(\mathbf{F_v}'\mathbf{F_v}) = \mathrm{tr}\,(\mathbf{F_v F_v}') = H_a$, and $\mathrm{tr}(\mathbf{G'G}) = \mathrm{tr}(\mathbf{GG'}) = H_d$. We further note that all orthogonal rotations of $\mathbf{F_v}$ will yield the same numerical value for $H_a$.

Similarly, all orthogonal rotations of $\mathbf{G}$ yield the same numerical value for $H_d$. Thus, if we obtain the roots and vectors of $(\mathbf{G'G})$ and write the equation $\mathbf{A_g}'(\mathbf{G'G})\mathbf{A_g} = \mathbf{D_g}$, the trace of $\mathbf{D_g}$ will still be equal to $H_d$. We could have also arrived at this conclusion by noting that the trace of a positive semi-definite matrix such as $\mathbf{G'G}$ is invariant under orthogonal rotation, and $\mathrm{tr}\,\mathbf{D_g}$ is equal to $H_d$.

Since the elements of $\mathbf{D_g}$ can be successively maximized and associated with a specific column of a new $\mathbf{G}$ based on $\mathbf{F_v}\, \mathbf{A_g}$ one can select columns of $\mathbf{G}\, \mathbf{A_g}$ that can maximize the magnitude of $H_d$ provided by a given number of columns of $\mathbf{G}$. The selection of the $k$ columns of $\mathbf{G}$ corresponding to the $k$ largest elements of $D_g$ is equivalent to selecting the $k$ factors in $\mathbf{F}_v\mathbf{A}_g$ that yield the largest value for $H_d$.

The factor solution $\mathbf{F}_d = \mathbf{F}_v\mathbf{A}_g$ will provide a successive column by column maximization of $H_d$, and when the number of columns (factors) is equal to the number of criterion variables, $m$, $\mathbf{F}_d\mathbf{F}_d' = \mathbf{C}_p$. If we wish to retain $k$ orthogonal factors, $k < m$, that maximize $H_d$, we need only select the $k$ columns (factors) of $\mathbf{F}_d$ corresponding to the $k$ largest eigen values of $\mathbf{G}'\mathbf{G}$. When $k < m$ the reproduced matrix $\mathbf{F}_d\mathbf{F}_d'$ becomes an approximation of $\mathbf{C}_p$, a much poorer approximation than is provided by the best $k$ factors of $\mathbf{F}_c\mathbf{F}_c'$. However, the $k$ factors of $\mathbf{F}_d$ that provide the largest factor contributions can provide more PCE than can the factors of $\mathbf{F}_c$ that provide the largest factor contributions. We would expect the $k$ largest factors of $\mathbf{F}_c$ to provide more PSE than the $k$ largest factors of $\mathbf{F}_d$.

When $\mathbf{F}_d$ and $\mathbf{F}_c$ are both $m$ by $m$ matrices, the values for $H_d$ and $H_a$ are equal regardless of which of these two solutions is utilized. However, when the $k < m$ factors corresponding to the $k$ largest eigen values of $\mathbf{D}_c$ or $\mathbf{D}_g$ are selected for further use, $H_d$ is larger for $\mathbf{F}_d$ and $Ha$ is larger for $\mathbf{F}_c$. $H_d$ can be formulated as $H_a - (\mathbf{1}'\mathbf{C}\mathbf{1})/m$. The term subtracted from $H_a$, $(\mathbf{1}'\mathbf{C}\mathbf{1})/m$, is easily shown to be equal to $m$ times the sum of the squared column means of $\mathbf{F}_d$, since $\mathbf{1}'\mathbf{C}\ \mathbf{1} = \mathbf{1}'\mathbf{F}_d\ (\mathbf{1}'\mathbf{F}_d)'$. We see that the single factor with the largest contribution to $H_d$ is one which has a comparatively large value for $H_a$, but also has a smaller mean factor loading (coefficient) than is found in the largest PC factor in the joint predictor-criterion space, the factor which maximizes $H_a$.

We now outline how to create classification efficient composites corresponding to factors for use in conjunction with an equal or larger number of job families. We suspect that the dimensionality of the joint predictor-criterion space for selection/classification systems of the the military services can justify the use of from three to seven factor based composites--depending on the adequacy of the future operational predictor battery. We begin by finding the $k$ orthogonal factors in the joint predictor-criterion space that maximize $H_d$. We then rotate these $k$ factors to simple structure of the job/criterion variables against oblique factors; i.e., $\mathbf{F}_d$ would be rotated and $\mathbf{T}_r$ obtained. This rotated solution, $\mathbf{F}_{dr} = \mathbf{F}_d\mathbf{T}_r$, can be extended back into the predictor space to obtain $\mathbf{F}_{tr}$ using the relationship $\mathbf{F}_{tr} = \mathbf{F}_t\mathbf{A}_g\mathbf{T}_r$. Note that the matrix $\mathbf{A}_g$ may have $k$ by $m$ columns and thus be an orthonormal rather than an orthogonal matrix, after the selection of k factors for rotation.

This process can be summarized in terms of the following supermatrices:

$$\mathbf{F} = \begin{bmatrix} \mathbf{RA_tD_t^{-1/2}\ A_g\ T_r} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \mathbf{VA_tD_t^{-1/2}\ A_g\ T_r} \end{bmatrix} = \begin{bmatrix} \mathbf{F_tA_gT_r} \\ \cdots\cdots\cdots \\ \mathbf{F_vA_gT_r} \end{bmatrix} = \begin{bmatrix} \mathbf{F_{tr}} \\ \cdots\cdots \\ \mathbf{F_{vr}} \end{bmatrix}.$$

The oblique factor solution, $\mathbf{F_{vr}}$, identifies factors that provide simple structure in the most classification efficient part of the job/criterion space. These factors are intuitively effective for classification purposes and can be precisely defined in terms of the predictor variables. While transformation to $\mathbf{F_d}$ provided the optimal space for classification (to the extent that $H_d$ is optimal), the rotation to simple structure in terms of the jobs provides a particular set of test composites (i.e., a set of rotated factor constructs) that has the potential for providing near optimal classification efficiency for assignment to a small number of job families.

After rotation in terms of the loadings of jobs on the factors, the solution is extended to test space for interpretation as to content. The extension of these classification efficient factors back to the predictor space can provide insight into the aptitudes measured by these factors.

Also, factor extension into test space serves to identify the test composites required to produce factor scores: $\mathbf{Q_r = Y\ R_t^{-1}F_{tr}}$. We propose these factor scores as the $k$ most classification efficient assignment variables that can be constructed on the basis of $\boldsymbol{H_d}$.

An optimal set of factor scores for use in a two stage selection-classification system, one in which FLS composites based on these factor scores are used to select and make assignments, would intuitively be based on a combination of factors from $\mathbf{F_c}$ and $\mathbf{F_d}$. One promising approach might base one factor variable on the largest factor of $\mathbf{F_c}$, $\mathbf{F_{c1}}$, and the remaining factor variables on classification efficient factors independent of $\mathbf{F_{c1}}$. This could be accomplished by first computing $(\mathbf{C_p - F_{c1}F_{c1}'})$ and then computing $\mathbf{G}$ and $\mathbf{F_d}$ in this residual space. Using the same notation as above, the rotated classification efficient solution in the criterion space is designated as $\mathbf{F_{vr}}$ and the solution for predictor variables against these same factors as $\mathbf{F_{tr}}$; we will refer to these solutions in residual space, respectively, as $\mathbf{F_{tr1}}$ and $\mathbf{F_{vr1}}$. The factor solution used to define factor scores in terms of predictor variables could thus be written as a super-matrix as follows: $\mathbf{F = (F_{tc1} \mid F_{tr2})}$.

# APPENDIX 3C
# ALGORITHM FOR SEQUENTIAL TEST SELECTION

## APPENDIX 3C.1: OVERVIEW OF APPROACH

The test selection algorithm described in this appendix has a separate module referred to as the figure of merit. We have described only one figure of merit in the context of the algorithm, the point distance index (PDI). This algorithm has been incorporated into a FORTRAN program with several alternative figures of merit, including: $H_d$, *Max-PSE*, and both $H_d$ and $H_a$ modified to avoid HC effects (all five indices are described in the text). This FORTRAN program has been applied to two data sets each with two subsets of 9 jobs and one subset of 18 jobs; all data subsets had 29 predictor variables from which to select.

The intercorrelation matrix among predictor tests, $\mathbf{R}_t$, can be factored by the square root (triangular) method in which each orthogonal factor is all, or part, of a predictor variable. This factor solution can be extended into the joint predictor-criterion space, yielding the factor matrix $\mathbf{F}$. Thus

$$\begin{bmatrix} \mathbf{R}_t \\ \cdots \\ \mathbf{V} \end{bmatrix} \underline{\text{factor solution}} \quad \mathbf{F} = \begin{bmatrix} \mathbf{F}_t \\ \cdots \\ \mathbf{F}_v \end{bmatrix} .$$

Building $\mathbf{F}_v$ , one factor at a time, with each factor consisting of an orthogonal component of a test corresponding to one "selected" test, a test is selected to maximize a function of $\mathbf{F}_v$. In our sequential test selection algorithm, the factoring process remains the same regardless of the function maximized; only the function of $\mathbf{F}_v$ changes to represent the different figures of merit (i.e., $H_d$, $H_a$, Max-PSE, PDI, etc.).

The algorithm for sequentially factoring the $n$ by $n$ intercorrelation matrix $\mathbf{R}_t$, and the extension of this solution to the $m$ by $m$ matrix $\mathbf{V}$, involves applying the same rule to each $\mathbf{F}_v$ type solution with $k$ columns ($\mathbf{F}_{vk}$) to produce $n-k$ $\mathbf{F}_v$ type matrices with $k$+1 columns in order to select the particular $\mathbf{F}_{v(k+1)}$ that maximizes the figure of merit. The application of this rule to $\mathbf{F}_{vk}$ is repeated until the desired number of tests are selected. In each repetition (iteration) $\mathbf{k} = \mathbf{k} + 1$. All tests once selected remain selected (although we know, strictly speaking, this is not optimal, i.e., the maximization of the figure of merit will only be approximated); we refer to the number of selected tests as $\mathbf{k}$.

At each step, $\mathbf{R}_t$ is conceptually divided into the $k$ variables that define the $k$ factors and the remaining $n-k$ test variables to which these factors are extended. In the factoring process there is no distinction between extending a factor solution to these $n-k$ remaining test variables (to produce $\mathbf{F}_{ek}$) or to the $m$ criterion variables to produce $\mathbf{F}_{vk}$. Thus, if we write the above transformation process for factoring $\mathbf{R}_t$ and extend this solution to $\mathbf{V}$, using further detail in our notation to reflect the distinction between predictor variables described above, we have the following relationship:

$$\begin{bmatrix} \mathbf{R}_t \\ \cdots \\ \mathbf{V} \end{bmatrix} \underline{\text{factor solution}} \begin{bmatrix} \mathbf{F}_{qk} \\ \cdots \\ \mathbf{F}_{ek} \\ \cdots \\ \mathbf{F}_{vk} \end{bmatrix} = \mathbf{F}_{sk};$$

$\mathbf{F}_{qk}$ will be a square matrix with all zeros above its diagonal elements (i.e., it is a triangular matrix), and $\mathbf{F}_e$ and $\mathbf{F}_v$ will be obtained by applying the same rule (multiplying by the same column vector) with respect to the $n - k$ remaining variables of $\mathbf{R}_t$ and the m variables of V respectively. Thus, $\mathbf{F}_{qk}$ is a k by k matrix, $\mathbf{F}_{ek}$ is a $n - k$ by k matrix, and $\mathbf{F}_{vk}$ is a m by k matrix.

We use the index $p$ to indicate the trial variables as they are being used as a candidate for selection as the next "best" test. Our algorithm calls for proceeding from $\mathbf{F}_{q(k-1)}$ to $\mathbf{F}_{qk}$ by bordering $\mathbf{F}_{q(k-1)}$ with a trial column of coefficients from $\mathbf{R}_t$ and $\mathbf{V}$, that is, $\mathbf{T}_{kp}$; $\mathbf{T}_{kp}$ is a selected column from

$$\begin{bmatrix} \mathbf{R} \\ \cdots \\ \mathbf{V} \end{bmatrix},$$

The selected column from this super matrix is used to bound $\mathbf{F}_{sk}$ to the right. $(\mathbf{F}_{S(k-1)} \mid \mathbf{T}_{kp})$ is multiplied by a $k + 1$ element column vector we will call $\mathbf{M}_{kp}$ to form a column vector of partial correlation coefficients, $r_{i(p.12)}$ when $k = 3$ indicating that a trial variable designated as $p$ in the above formula will become variable # 3 if selected; we will call this column of partial correlation coefficients $\mathbf{H}_{kp}$. Thus,

$$\begin{bmatrix} F_{q(k-1)} & \vdots & T_{qkp} \\ \cdots & \vdots & \cdots \\ F_{e(k-1)} & \vdots & T_{ekp} \\ \cdots & \vdots & \cdots \\ F_{v(k-1)} & \vdots & T_{vkp} \end{bmatrix} M_{kp} = H_{kp} \,; \qquad \begin{bmatrix} F_{q(k-1)} \\ \cdots \\ F_{e(k-1)} \\ \cdots \\ F_{v(k-1)} \end{bmatrix} = F_{s(k-1)p} \quad ,$$

and $(\mathbf{F}_{s(k-1)} \mid \mathbf{H}_{sk}) = \mathbf{F}_{sk}$ ; $\mathbf{H}_{vk}$ being the best of all the $\mathbf{H}_{vkp}$ vectors, and $\mathbf{H}_{vk}$ is a subset of $\mathbf{H}_{sk}$.

$\mathbf{T}_{qkp}$ is a null vector, all elements are equal to zero. The value of $(1 - R_p^2)$ is substituted for 1.0, i.e., the element which has the value of 1.0 is instead given the value of $(1 - R_p^2)$. $R_p^2$ is equal to the sum of squares of all elements in a row that are to the left of the diagonal element in $\mathbf{F}_e$. If selected, this $p^{th}$ variable will become the next row and column of $\mathbf{F}_q$. Thus $\mathbf{F}_q$, always a square matrix has a 1.0 as its first diagonal element and $(1 - R_p^2)^{1/2}$ as the value of each succeeding diagonal element. The sums of squares of each row of $\mathbf{F}_q$ is always equal to 1.0.

A matrix $\mathbf{F}_{kp}$ must be computed for each of the $n-k$ unselected tests (the tests in the $e$ set) and the figure of merit , $f_m(\mathbf{F}_{vkp})$, with $p$ taking values for $p = k$ to $n$. The test yielding the largest figure of merit is selected and becomes the next test to be taken from $\mathbf{F}_e$ and placed in $\mathbf{F}_q$. The column vector $\mathbf{M}_{kp}$ is derived from the row vector of $\mathbf{F}_{ek}$ representing the same test variable as is represented by the column vector $\mathbf{T}_{kp}$. For example, for a $i^{th}$ row of $\mathbf{F}_{ek}$ corresponding to $\mathbf{T}_{kp}$ (let $k = 3$ in our following example), the elements of this row vector would be partial correlation coefficients as follows: $r_{i1}$, $r_{i(2.1)}$. The three elements of $\mathbf{M}_{2p}$ would be as follows: $r_{ip}/(1-R_p^2)^{1/2}$, $r_{i(2.p)}/(1-R_p^2)^{1/2}$, $1/(1-R_p^2)^{1/2}$. The value of $R_p^2$ in our example is equal to the sum of $(r_{p1})^2$ and $(r_{p2.1)})^2$. This $R_p^2$ is obviously the squared multiple correlation coefficient of the $p^{th}$ test variable with test variables 1 and 2.

In the next step the best of the trial variables is designated as variable 3 and $r_{i(3.12)}$ is computed for all the remaining variables (i.e., for all variables other than 1, 2, and 3). This next step is a factor extension process, from $\mathbf{F}_{qk}$ to $\mathbf{F}_{ek}$ and $\mathbf{F}_{vk}$. In this factor extension process $\mathbf{M}_2$ is based on $R_k^2 = (r_{13})^2 + (r_{3(2.1)})^2$, in contrast to the definition of $R_p^2$ as equal to $(r_{p1})^2 + (r_{p(2.1)})^2$

We will define the more general algorithm after first providing an example with a specified figure of merit and $k$ successively taking values of 1, 2, and 3. We use PDI as the figure of merit and apply rules for selecting the first, second and third tests to approximate a maximization of PDI using a sequential process in which each "best" test is retained in the battery without further question once it is selected. This process can be easily extended to the selection of four or more tests.

## APPENDIX 3C.2: PDI EXAMPLE; SELECTING THE FIRST AND SECOND TEST

Prior to selecting the first test $k$ is equal to zero and $\mathbf{F}_{sk}$ is a null set. Thus , our figure of merit must be computed directly on each $\mathbf{F}_{s1p}$. For our PDI example this calls for summing the absolute values of the differences from the column means for each trial $\mathbf{T}_{v1p}$ (i.e., a column of $\mathbf{V}$). The test which yields the largest sum is selected as the first factor. We designate this first selected test as test variable 1, and the column of $\mathbf{R}$ bordered below by $\mathbf{V}$ corresponding to this best test is identified as $\mathbf{T}_{s1}$; $\mathbf{F}_{sk} = \mathbf{T}_{sk}$ only when $k = 1$.

We now have a $\mathbf{F}_{sk}$ ($k = 1$) a column vector with $n+m$ elements, which can be bordered with $\mathbf{T}_{2p}$ in order to commence the process of selecting the second test. $\mathbf{M}_{2p}$ has its two elements as follows: $-r_{ip}/(1-(r_{i1})^2)^{1/2}$, $1/(1-(r_{i1})^2)^{1/2}$. We now border $\mathbf{F}_{s1}$ with $\mathbf{T}_{s2p}$ and compute (assemble) the $(n+ m)$ by 2 matrix $(\mathbf{F}_{s1} \mid \mathbf{T}_{s2p})$. From this matrix, $\mathbf{F}_{s1p}$ bordered by $\mathbf{T}_{s2p}$, each $\mathbf{M}_{2p}$ vector, as a function of the $p^{th}$ row of $\mathbf{F}_{e1p}$, can be computed. Using the column vector $\mathbf{H}_{s2p}$, where $\mathbf{H}_{s2p} = (\mathbf{F}_{s1} \mid \mathbf{T}_{s2p})\, \mathbf{M}_{2p}$, $\mathbf{F}_{s2p} = (\mathbf{F}_{s1} \mid \mathbf{H}_{s2p})$.

A figure of merit is computed for each trial $F_{v2p}$ and the predictor test corresponding to the largest PDI selected and designated as test variable 2; The matrix $\mathbf{F}_{v2p}$ is obtained using the same process (using the same $\mathbf{M}_{2p}$) as produces $\mathbf{F}_{s2}$ as described above. $\mathbf{H}_{v2p} = (\mathbf{F}_{v1} \mid \mathbf{T}_{v2p})\, \mathbf{M}_{2p}$ ,and $\mathbf{F}_{v2p} = (\mathbf{F}_{v1} \mid \mathbf{H}_{v2p})$. Once the value of $p$ associated with the $\mathbf{T}_{v2p}$ which yields the $\mathbf{H}_{v2p}$ which in turn provides the best figure of merit has been identified, we designate the $\mathbf{M}_{2p}$ associated with the best $\mathbf{T}_{e2p}$ as simply $\mathbf{M}_2$ and define the corresponding best test as test 2. The column of $\mathbf{R}$ bounded below by $\mathbf{V}$ corresponding to test 2 is now designated as $\mathbf{T}_2$. We will now wish to compute $\mathbf{H}_{s2}$ as a function of $\mathbf{M}_2$, $\mathbf{F}_{s1}$ and $\mathbf{T}_{s2}$. $\mathbf{F}_{s2}$ can be computed as follows: $\mathbf{F}_{s2} = (\mathbf{F}_{s1} \mid \mathbf{H}_{s2})$, $\mathbf{H}_{s2} = (\mathbf{F}_{s1} \mid \mathbf{T}_{s2})\, \mathbf{M}_2$ , where $\mathbf{M}_2$ is equal to the column vector,$(-r_{1p}/(1-(r_{p1})^2)^{1/2}$, $1/(1-(r_{p1})^2)^{1/2})$, and $\mathbf{T}_2$ is equal to the column vector: $(r_{12}, 1.0, r_{32}, \ldots r_{n2}, \ldots v_{12}, \ldots v_{m2})$,where the second best test is designated as variable 2. Note that $\mathbf{T}_k$, in general, is equal to the column vector: $(r_{1k}, r_{2k}, \ldots r_{nk}, \ldots v_{1k}, \ldots v_{mk})$.

The figure of merit for selecting the test to be designated as test 2 is, in our example, PDI. We compute PDI from each trial $\mathbf{F}_{vkp}$, $\mathbf{F}_{vkp} = (a_{ij})$, as

$PDI = SUM_i^m(SUM_j^k(a_{ij} - a_j^*)^2)^{1/2}$. The test that yields the largest value of PDI is designated as test 2. Note that $a_j^*$ is the mean of the m values of $a_j$ for the $j^{th}$ column of $\mathbf{F}_v$.

## APPENDIX 3C.3: PDI EXAMPLE; SELECTING THE THIRD TEST

The same process as is used to select test 2 is used to select test 3. We commence with the $(n + m)$ by 3 matrix ,$(\mathbf{F}_{s2} \mid \mathbf{T}_{s3p})$ and compute $\mathbf{M}_{3p}$ as a function of the $p^{th}$ row of $\mathbf{F}_{e2}$; the two elements of this $p^{th}$ row are squared and summed to provide $R_p{}^2$ in the computation of the $\mathbf{M}_{3p}$ to be used in conjunction with $\mathbf{T}_{v3p}$. The first two elements of the $p^{th}$ row are multiplied by $- 1/(1-R_p{}^2)^{1/2}$ to provide the first and second elements of $\mathbf{M}_{3p}$; the third element of $\mathbf{M}_{3p}$ is equal to $1/(1-R_p{}^2)^{1/2}$. Thus $\mathbf{M}_{3p}$ is a column vector as follows: $(-r_{p1}/(1-R_p{}^2)^{1/2}, -r_{p(2.1)}/(1-R_p{}^2)^{1/2}, 1/(1-R_p{}^2)^{1/2})$. $R_p{}^2 = (r_{p1})^2 + (r_{p(2.1)})^2$.

The column vector , $\mathbf{T}_{kp}$, and the row vector used to compute $\mathbf{M}_{kp}$ represent the same test variable. While $\mathbf{T}_{kp}$ consists of correlation coefficients ($p^{th}$ column of the super matrix $\mathbf{R}_t$ bordered below by $\mathbf{V}$), the $p^{th}$ row vector of $\mathbf{F}_{e(k-1)}$ used to compute $\mathbf{M}_{kp}$ is of course made up of factor coefficients.

The best $\mathbf{T}_{kp}$ is the one which provides the best figure of merit resulting from the use of that $\mathbf{T}_{kp}$, with $p$ taking on $n-k$ values to represent the remaining unselected tests. The selection of the best $\mathbf{T}_{kp}$ provides for the identification of the $k^{th}$ "best" test. When we have just selected the third test ($k = 3$), our next step is to extend the solution $\mathbf{F}_{qk}$ to create $\mathbf{F}_{ek}$ and $\mathbf{F}_{vk}$. In this factor extension process $\mathbf{M}_3$ is based on $(R_k)^2 = (r_{13})^2 + (r_{3(2.1)})^2$, when $k = 3$.

Each iteration in which one more test is selected requires the computation of $n$-$k$ column vectors, each having $k$ elements, to be used as a trial multiplier of each row in $(\mathbf{F}_{v(k-1)} \mid \mathbf{T}_{vkp})$ to produce the $k^{th}$ column of $\mathbf{F}_{vkp}$ that in turn produces the largest PDI value. Each of the first $k-1$ elements of $\mathbf{M}_{kp}$ ( remember that this column vector has $k$ elements) is equal to $a_{pj}(-1/(1-R_p{}^2)^{1/2})$, where $R_p{}^2 = SUM_j^k\ (r_{pj})^2$ and $a_{pj}$ is a factor coefficient in the $p^{th}$ row and $j^{th}$ column of $\mathbf{F}_{e(k-1)}$; the last element of $\mathbf{M}_{kp}$ is equal to $1/(1-R_p{}^2)^{1/2}$. All cells above the diagonals in $\mathbf{F}_{qk}$ will always be equal to zero.

## APPENDIX 3C.4: DERIVATION OF FORMULAE FOR $\mathbf{M}_{kp}$ AND $\mathbf{M}_k$

The creation of a trial column, $\mathbf{H}_{vkp}$, to border $\mathbf{F}_{v(k-1)}$ to permit the selection of the $k^{th}$ best test requires the computation of $\mathbf{M}_{kp}$. When the best $(\mathbf{F}_{v(k-1)} \mid \mathbf{H}_{vkp})$ has been selected, the corresponding $\mathbf{T}_{kp}$ and $\mathbf{M}_{kp}$ have also been selected, and $\mathbf{M}_k$ has been defined. For $k = 3$, we express $\mathbf{F}_{s(k-1)}$ in terms of the $i^{th}$ row vector of this factor matrix and similarly express $\mathbf{H}_{sk}$ and $\mathbf{T}_{sk}$ in terms of correlation coefficients and demonstrate that $(\mathbf{F}_{s(k-1)} \mid \mathbf{T}_k)\,\mathbf{M}_k$, does indeed equal $\mathbf{H}_k$.

We first define $\mathbf{H}_k$ in terms of its $i^{th}$ row element, $r_{i(3.12)}$. This is the correlation of the $i^{th}$ variable with the component of variable 3 that is orthogonal to variables 1 and 2. It can be readily shown that $r_{i(3.12)} = (r_{i3} - r_{i1}\, r_{13} - r_{3(2.1)}\, r_{i(2.1)})/(1-R_k{}^2)^{1/2}$. Using this same notation, the $i^{th}$ row of $(\mathbf{F}_{s(k-1)} \mid \mathbf{T}_k)$ can be written as follows: $(r_{i1}, r_{i(2.1)}, r_{i3})$. For $k = 3$ our definition of $\mathbf{M}_k$ provides the following column vector: $(-r_{31}/(1-R^2)^{1/2}$, $-r_{3(2.1)}/(1-R^2)^{1/2}$, $1/(1-R^2)^{1/2}$ ). Using this notation it is clear that $(\mathbf{F}_{s(k-1)} \mid \mathbf{T}_k)\,\mathbf{M}_k = \mathbf{H}_k$, when $k = 3$. It is also easy, although not accomplished here, to prove the general case, i.e., for $k$ equal to any value.

# CHAPTER 4. MODEL SAMPLING AND SIMULATION AS A TOOL FOR MEASURING UTILITY

## A. INTRODUCTION

A simulation capability which can provide accurate, defensible estimates of mean predicted performance (MPP) as the outcome of any prescribed assignment process without the need to make questionable assumptions, while precisely reflecting a defined applicant population, is essential to the credible estimation of the utility of selection/classification practices. An adequate simulation approach should permit the determination of MPP for both the theoretically optimal and the invariably flawed operational assignment processes.

The relatively simple analytical techniques useful in computing MPP in the selection mode are not similarly useful in the classification mode. Although the required means and variances of predicted performace for selected and allocated groups can be defined in terms of definite multiple integrals, integration of the required functions of the multivariate normal distribution produces mathematical equations too complicated for practical use.

Many classification problems can be expressed in terms of definite integrals of the normal multivariate distribution, defining assignment regions by half-hyperplanes (see Lord, 1952). In a paper presented at the 1985 National APA Convention (McLaughlin, Rossmeissl, Wise, Brandt, and Wang, 1985), the author concurs with Lord's statement that "the necessary expressions at present available for the integrals are too cumbersome to be of practical use" (Lord, 1952). This statement is just as true today as when Lord published his article.

Such problems can be solved by a model sampling approach. Model sampling (Johnson and Sorenson, 1974) provides not only a practical way to solve such mathematical equations but also the flexibility to impose operational procedures and conditions precisely on the assignment problem.

Our primary concern in this chapter is how to obtain MPP scores as measures of PUE, PSE or PCE. However, this approach can also be used to determine whether adequate numbers of qualified incumbents will result from specified cut scores or training policies, or how many more applicants would have to be recruited if minimum qualifications were to be raised. The possibilities for using model sampling in experiments with simulated systems are almost inexhaustible.

The term model sampling implies the generation of synthetic scores that have statistically equivalent properties as contrasted to empirical scores. That is, synthetic scores generated to have the characteristics of test scores in a predictor battery would yield the same covariance matrix as would the empirical scores, provided both samples were sufficiently large. The covariance matrices for a number of synthetic score samples would vary around the covariance matrix selected to represent the universe, much as would covariance matrices based on samples of empirical scores drawn from a universe of score vectors.

When large data bases containing test scores and values of other relevant variables exist (several years of Army input are available for use as a result of Project A), such data can frequently be used instead of synthetic scores. Both pros and cons to the use of such simulations in the place of model sampling should be considered. The shape of the score distribution, with all its warts and blemishes, will be more realistic for a simulation using empirical scores as compared to synthetic scores generated to have a normal distribution. However, with a little extra effort, synthetic scores can be generated to reflect any degree of censoring that is desired, and thus produce distributions closer to a distribution of a future population than is provided by the detailed shape of the distributions of the past years.

Model sampling has increased flexibility over simulations using data base scores. Samples of any number and size can be generated for any universe, including a current or future youth population, if that universe can be defined by both the covariances among the relevant predictor variables and the validities of these variables against all criterion components. Meeting these conditions permits the exploration of selection policies that would produce a different input than is present in the data bank, and permits a more direct basing of selection/classification results on a youth population than is possible using data base scores. Also, empirical groups of soldiers possessing certain scores are sometimes small in number and the resulting necessity to use incomplete data may produce empirical correlation matrices that are not positive semi-definite (i.e., could not occur as the result of analyzing complete real data sets).

## B. MODEL SAMPLING CONCEPTS

### 1. Generating Synthetic Scores with Designated Expected Covariances

Simulations to determine the PUE, PSE, or PCE of alternative sets of predictors, selection/assignment processes, criteria, or job structures can use scores either from data banks or from the generation of synthetic scores. A vector of synthetic scores representing an artificial person, or "entity," should have the same statistical properties as random samples of empirical scores drawn from the relevant universe of such scores.

We will consider synthetic scores to be adequate for the simulation of a system that includes personnel procedures. We will also consider the impact that personnel decisions have on job performance provided the scores have the desired Gaussian shape to their distribution and also have their expected means and covariance matrices equal to the universe values. This universe should represent the personnel entering the system, the youth population in general, applicants, trainees, workers eligible for the first stage in the system being simulated. In the remainder of this section we will assume that: (1) we have universe covariance matrices representing the desired universe, and (2) both predictors and performance measures are appropriately depicted as having normal (Gaussian) distributions.

We will later discuss the mechanics of how to generate an $N$ by $n$ matrix of normal deviates, $\mathbf{X}_n$, such that the expected matrix, $1/N\ \mathrm{E}(\mathbf{X}_n'\mathbf{X}_n)$, is equal to an $n$ by $n$ identity matrix. We designate a "score" matrix in which each element is divided by the square root of $N$ by writing the matrix in caps, bold face and underlined, thus, $1/N\ \mathrm{E}(X_n'X_n) = I_n$. The test scores we wish to generate, in standard score form, are considered in samples of $N$ entities as $N$ by $n$ matrices referred to as $\mathbf{Y}$. Where $\mathbf{R}_t$ is the matrix of correlation coefficients among the tests in the universe, $\mathrm{E}\ 1/N(\mathbf{Y}'\mathbf{Y}) = \mathbf{R}_t$. Similarly, for a set of $m$ jobs predicted by these $n$ tests, the $N$ by $m$ matrix of predicted performance scores (LSEs) is designated as $\mathbf{Z}$, and $1/N\ \mathrm{E}(\underline{\mathbf{Z}}'\underline{\mathbf{Z}}) = \mathbf{C}$; $\mathbf{C}$ is the matrix of universe values for the covariances among the predicted performance measures. We will show that we have the option of generating the $\mathbf{Y}$ matrix as a transformation of $\mathbf{X}_n$, and using a regression equation to compute the scores in the $\mathbf{Z}$ matrix, or, alternatively, of directly generating $\mathbf{Z}$ as a transformation of an N by m matrix of normal deviates, $\mathbf{X}_m$, whichever is more economical (depending on whether $m$ or $n$ is larger), or whichever best suits the research design.

The matrix $\mathbf{Y}$ can be generated from $\mathbf{X}_n$ using a transposed factor solution of $\mathbf{R}_t$ as the transformation matrix. That is, $\mathbf{X}_n\mathbf{F}_t' = \mathbf{Y}$, and $E\ 1/N(\mathbf{Y}'\mathbf{Y}) = \mathbf{R}_t$ where $\mathbf{F}_t\mathbf{F}_t' = R_t$. This relationship becomes apparent from observing that $\mathbf{Y}'\mathbf{Y} = \mathbf{F}_t\mathbf{X}_n'\mathbf{X}_n\mathbf{F}_t'$, and that $E(\mathbf{F}_t\mathbf{X}_n'\mathbf{X}_n\mathbf{F}_t') = \mathbf{F}_tE(\mathbf{X}_n'\mathbf{X}_n)\mathbf{F}_t' = \mathbf{F}_t\mathbf{I}\mathbf{F}_t' = \mathbf{R}_t$. Similarly, $\mathbf{X}_m\mathbf{F}' = \mathbf{Z}$, where $\mathbf{F}\mathbf{F}' = \mathbf{C}$, and $E(\mathbf{Z}'\mathbf{Z}) = \mathbf{F}(E(\mathbf{X}_m'\mathbf{X}_m))\mathbf{F}' = \mathbf{F}(\mathbf{I}_m)\mathbf{F}' = \mathbf{C}$.

### 2. Synthetic Factor Scores

Synthetic factor scores with expected means of zero, standard deviations of 1, and expected intercorrelations of zero can be readily produced. A sample of $N$ entities, each with $k$ factor scores, provides a matrix $\mathbf{X}_k$, where $E(\mathbf{X}_k'\mathbf{X}_k) = \mathbf{I}_k$. Both predictor and criterion variables can be expressed in terms of these factor scores.

When the columns of a factor solution represent the hypothetical constructs commonly referred to as factors, the rows provide the regression weights which, when applied to the factor scores, produce an LSE of the variable represented by the row. If the factor solution reproduces a correlation matrix with communalities in the diagonals, $\mathbf{F}_h\mathbf{F}_h' = \mathbf{R}_h$, the matrix of scores produced by applying the "best" weights, $\mathbf{F}_h$, to the factor scores in $\mathbf{X}_k$ (i.e., $\mathbf{Y}_h = \mathbf{X}_k\mathbf{F}_h'$) represents the row variables in common space. However, if ones are placed in the diagonals of the correlation matrix and the factorization is complete, scores for the row variables will be provided in total space and $\mathbf{X}_K\mathbf{F}_t' = \mathbf{Y}$ while $\mathbf{X}_k\mathbf{F}' = \mathbf{Z}$; $k$ denotes the number of columns in the factor solution and will be equal to $n$ or $m$, respectively, when a complete factorization is accomplished in the "total" space.

## C. EARLY USE OF MODEL SAMPLING IN PSYCHOLOGICAL RESEARCH

Lewis (1975) provides a brief description of the historical background of the sampling process we call model sampling. Sampling is said to have been first used by "student" to determine the $t$ distribution. Student's population was obtained by selecting 3,000 pairs of index finger measurements of criminals; these measurements were written on cards and sampled by drawing from a shuffled deck.

Bispham apparently was the first to sample from an arbitrary theoretical population. A population of 30 counters was drawn, without replacement, from urns.

The first published tables of random digits, attributed to Tippit, were sampled from 1,000 small cards placed in a bag. After each digit was recorded, the card was replaced and the cards in the bag were mixed well. The numbers drawn from the bag were converted into random numbers uniformly distributed through use of a key. Finally, in 1951, the Rand corporation produced one million random digits by using a mechanical device; the tools for accomplishing the first step in model sampling--the generation of random (or pseudorandom) numbers--became generally available to researchers.

Model sampling differs from Monte Carlo techniques in that the latter does not necessarily involve the simulation of a decision process or the behavior of a system (the integration of such processes) but does, whenever possible, use variance reduction techniques that provide a given level of fidelity with fewer observations. Most such variance reduction "tricks" further remove the similarity of the Monte Carlo process to any real process or system. Model sampling is a means of accomplishing simulations just as Monte Carlo techniques are numerical analysis tools for the solution of mathematical problems (Wagner, 1969). Model sampling is a "technique of abstracting a system in terms of the statistical properties of its entities and the operations to be performed on those entities." (Johnson and Sorenson, 1974, p. 38).

Kaiser and Dickman (1962) provide a method of generating synthetic scores "to yield sample **R**'s from an arbitrary population **R**" (p. 179); their method is essentially the same as is described in the previous section. The authors use a simplex correlation matrix given by Guttman as the **R** in their example. They first computed a principal component (pc) factor solution of **R**, generated a sample of scores, computed a pc solution of **R**. They then used this second transformation matrix on the original sample of random normal deviates to generate scores yielding a correlation matrix equal to the population matrix **R**.

Wherry, Naylor, Wherry, and Fallis (1965) review the literature on "integrating random error into known functions to generate fictitious data" (p. 304) whereby to test methods of fitting fallible data. They describe their own method of generating synthetic trait scores when providing stimuli for raters in an Air Force experiment. Their method differs from that of Kaiser and Dickman in that Wherry et al. commence their procedure from a population factor structure instead of from a population correlation matrix that requires factoring to obtain a transformation matrix. The use of their method "to test cross-validity of different methods of test selection and prediction" (p. 311), as well as the conduct of a variety of experiments involving the evaluation of profiles by raters are proposed.

Three methods for generating multivariate normal samples of synthetic scores from a population with prescribed intercorrelations are compared by Barr and Slezak (1972). All three methods require the generation of random normal deviates and their transformation into scores having the desired expected covariances. One method is called the rotation method and is based on the existence of a "matrix **P** such that $\mathbf{P'CP} = \mathbf{I}$' (p. 1049); the transformation matrix would thus be equal to the pc factor solution of **C** (to $F_d'$ or $\mathbf{D}_d{}^{1/2}\mathbf{A}'$ in the notation of Chapter 3). The second method is referred to as the "conditional method" and the third and "best" method is the "triangular factorization method." "Best" is defined in terms of the time required to compute the transformations and to generate 1,000 score vectors. There are, of course, issues to consider other than the computational economy associated with computing a score vector (e.g., the amount of information pertinent to PCE provided by a given number of factors of a given type).

A more detailed description of the generation of random scores having a multivariate normal distribution is provided by Naylor, Balintify, Burdick, and Chy (1966). These authors recommend (pp. 97-99) use of a "square root method" factor matrix (another name for the triangular factorization mentioned above) as the multiplier of vectors of normal variates with zero mean and unit variance. Most of their description is taken up with an explanation of the square root factoring method. They apparently consider sufficient justification for this approach to be the probability density function of **X** to be: $f(x) = [\det (2 \text{ pi } \mathbf{R})]^{-1/2} \exp [-1/2 \, (\mathbf{X'R}^{-1}\mathbf{X})]$.

The model sampling approach reported by Sorenson (1965a, 1965b), Niehl and Sorenson (1968), Olson, Sorenson, Haynam, Witt, and Abbé (1969), Johnson and Sorenson (1974), and by others of the same Army research team, is among the first in the literature that report MPP standard scores as the outcomes of simulations of personnel selection/classification procedures.[19] A number of design issues arose in these early applications of the model sampling approach to the evaluation of PUE or PCE--issues that did not have to be faced in the psychometric studies discussed in the paragraphs above. Some of these issues pertaining to the simulation of personnel systems will be discussed in the following section.

---

19 Sorenson's doctoral dissertation in 1965, University of Washington, used MPP standard scores as outcomes in a simulation that utilized empirical data; Brogden (1946b) pioneered the concept of equating PCE to MPP but did not make use of model sampling.

## D. GENERATING SYNTHETIC SCORES FOR MODEL SAMPLING

### 1. Pseudorandom Numbers

A pseudorandom number generator is a program to simulate a sample drawn from a population with known distribution characteristics. Such programs produce a repeatable finite sequence of numbers which can be perfectly predicted from the initial conditions [algorithm, parameter(s), and seed]. The aim is for the sequence to possess the essential statistical properties of a truly random sequence, so that it can be justifiably used in place of a random sample drawn from a specified universe.

The construction and testing of random number generators by themselves constitute a major field of study. Would-be developers of their own generators or even of their own parameters should consult the extensive literature on this topic. Here we will discuss this topic just enough to afford the reader an opportunity to become an educated consumer. Most scientific, statistical or simulation software packages have at least one built-in generator (or, more usefully, a subroutine) whereby to produce uniformly distributed (rectangular) random numbers; many will have generators of Gaussian distributed numbers, and some will have generators that yield numbers with Gamma, Beta, or Poisson distributions.

Caution must be exercised in the use of some of these readily available generators; a few are suitable only for recreational games or classroom exercises. Any scientist should want to have documentation on the generator being considered for use in conducting a model sampling experiment, and insist on providing one's own carefully recorded seed to assure that the experiment can be truly replicated.

Researchers planning to use a readily available random number generator will either commence with uniformly distributed numbers to be transformed into a distribution of another shape (usually Gaussian), or will use a generator that directly produces numbers with the desired distribution. All researchers can take precautions that will reduce negative effects that the use of a questionable generator in their study can have on the credibility of their results. Research strategies to minimize the impact that unwanted regularities and other defects in the generator can have on model sampling results will be discussed later.

## 2. The Rectangular Distribution

The uniform, rectangular distribution of numbers for a finite set ranging from zero to one has a mean of 0.5, and a variance of (1/12) or 0.08333, a third central moment of zero, and a fourth central moment of 0.0125. The output of a random uniform number generator should approximate these values for the first four central moments. Alternative generators can be compared with respect to how closely these theoretical values are met for the sample sizes to be used in an experiment. More precise goodness-of-fit tests can be conducted by dividing the zero-one interval into classes and computing Chi-square. Similarly, the Kolmogorov-Smirnov test can be performed for the accumulative step function over the zero-one interval.

A satisfactory random number generator must provide more than a good fit to the desired theoretical distribution; it must also exhibit apparent independence among the numbers output by the generator and must not exhibit unwanted regularities or patterns. Marsaglia (1968) warned that random numbers produced by multiplicative generators, when considered as coordinates of points located on a unit $n$-dimensional cube, will fall on a relatively small number of parallel hyperplanes, indicating that no single generator should be used to generate more than one element in an entity vector (one element in each row of the matrix $\mathbf{X}$).

The desired independence in generator output is frequently measured in terms of serial correlation, runs and the distributions of sums or maximum values in subsets. Dependence and regularities in the output of generators is appropriately more feared by investigators than are discrepancies from the shapes of theoretical distributions.

Most, if not all, pseudorandom number generators must be empirically tested to determine their suitability, and there is no theoretical basis for extrapolating from tested sequences to other untested sequences for the same generator. Also, most algorithms for pseudorandom number generators require different parameters for different sized computer words. Thus a sequence of random numbers created for a model sampling experiment cannot be replicated, with only a few exceptions, across different types of computers (e.g., different word length or ones versus twos complement machines). In general, the exact replication of a model sampling experiment can be accomplished only when the computer to which the experiment is being migrated is of the same word length and logical type.

Tausworthe (1965) provides a generator based on a theory which makes it independent of word length and which predicts good statistical behavior prior to empirical tests. Canavos (1967) makes an empirical comparison between a generator identified as GETRAN which implements Tausworthe's method and RANF, a generator based on the comm nly used congruence algorithm using parameters provided by Control Data for use on their 6,000 series computers. Thus RANF is the manufacturer's recommended uniform random number generator for the computer used to make the comparison.

Results for the two generators were similar except that GETRAN fared better on the tests for runs and for serial correlation when smaller sample sizes were used. "Based on these tests, the indication is that the random properties of the sequence generated by RANF decay as the sample size decreases" (p. 490). This is in marked contrast to GETRAN which gave good results for those smaller samples (e.g., $N$ = 200). The author recommends the use of Tausworthe's approach because: (1) it is machine independent, (2) it tests well, and (3) it is easily programmed (and executed) in FORTRAN without sacrificing any of its characteristics.

Those contemplating the conduct of a model sampling experiment may wish to learn more about uniform random number generators. One of the best introductions to pseudo-random number generators is provided by Knuth (1981). Knuth is very pessimistic regarding the quality of the readily available generators, warning that: "the most common generator in actual use, RANDU,[20] is really horrible" (p. 173). He urges students, in exercise 6, "Look at the subroutine library of each computer installation in your organization and replace the random number generators by good ones. Try to avoid being too shocked at what you find" (p. 176). McLaren and Marsaglia (1965) propose a number of tests for generators that go beyond the serial tests and chi-square tests for goodness of fit to the theoretical distributions (tests customary in 1965). Based on their own test results the authors propose combining generators.

We believe most researchers would prefer to use off-the-shelf software, but could readily provide their own generator. Those interested in generators for IBM-type computers should read the documentation for SSP. As noted above, RANDU, the generator in SSP, should not be considered adequate for more than preliminary analyses, student demonstrations, or games. In general, caution should be exercised in the use of generators provided by the computer center and should not necessarily settle for the generator built into a statistical package. Two very readable publications on pseudo-random number generators (Larkin, 1965; Kuehl, 1969) are provided by the Army research

[20] RANDU is the algorithm implemented in SSP.

team which first applied model sampling techniques to the evaluation of selection and classification strategies. Anyone with a need to provide themselves with an immediately acceptable generator should follow the advice of Park and Miller (1988).

Noting the frequent criticism of earlier pseudorandom number generators, and having an immediate need for a series of model sampling experiments conducted in 1989 through 1990, we designed our own generator. We concluded that we could appropriately indulge in overkill, providing ourselves with security against future criticism with very little cost in computer time, by using a number of independent generators to produce each score. As suggested by Park and Miller (p. 1197), we obtained the list of 205 "best" multipliers provided by Fishman and Moore (1986) and used these multipliers in separate generators for each variable.

## 3. Gaussian Distributions

The central limit theorem guarantees asymptotic normality of sums of independent random numbers regardless of the distribution of the individual numbers. The sum of $k$ independent variables, each with a standard deviation equal to $S$, will have a standard deviation of $(kS^2)^{1/2}$. Thus the sum of 12 uniform random numbers uniformly distributed over the interval of zero to one, each with $S^2 = 1/12$, will approximate a Gaussian distribution with a mean of 6 and a standard deviation of one over the range of zero to twelve. A Gaussian distribution should also approximate the equality of third and fifth moments to zero, the fourth moment to 3, and the sixth to 15. The formula for transforming a sum of uniform random numbers to a normal deviate with a standard deviation of one is as follows:

$$x_{ij} = [(\textstyle\sum_i^k u_i\ ) - k/2]\ (12/k)^{1/2} \quad ; \qquad (10.1)$$

where $u_i$ is the $i^{th}$ uniform random number of a series of such numbers going from 1 to k and $x_{ij}$ is the element of the $i^{th}$ row and $j^{th}$ column of the matrix $\mathbf{X}$, $\mathbf{X} = (x_{ij})$.

The central limit theorem also applies to the sum of random normal deviates. An unweighted, or weighted, sum of scores with an approximately Gaussian distribution will approximate the theoretical distribution more closely than do the individual scores. This incidental improvement resulting from the transformation process that provides the desired covariances is greater when the number of variables is larger and the average intercorrelation coefficient is smaller. Some improvement will always result from summing

$k$ scores and dividing by $(k)^{1/2}$. However, the researcher has more reason to be concerned with the generators output with respect to serial correlation, runs, and unwanted regularities (patterns) than with the closeness of the fit to the Gaussian distribution.

The selection/assignment process being simulated in a model sampling experiment may utilize predictors whose distributions are definitely not Gaussian. For example, the AFQT is expressed in percentile scores provided by a conversion of the test composite scores into a rectangular distribution. From the above discussion it is obvious that when a procedure in which synthetic scores with the required mean, standard deviation, and rectangular distribution are first created and subsequently transformed, along with other variables, to achieve the desired intercorrelations, the once rectangular variable will be found to have assumed an unwanted bell shape. The effect of the phenomenon known as the central limit theorem will also distort other non-normal distributions when the linear transformation for producing the desired $\mathbf{R}_t$ is applied.

A workable approach to achieve both the desired non-Gaussian distribution and the relationship $E(\underline{Y}'\underline{Y}) = \mathbf{R}_t$, is to first create normally distributed variables with expected covariances offset just enough to provide for the later distortion which will occur when selected variables are transformed into desired non-normal distributions. Boldt (1965) derives formulae for computing the amount of change occurring in the values of product moment correlation coefficients when the shape of one or both variables are changed from normal to rectangular, or vice versa. Formulae provided by Boldt show that the synthetic normal deviate that is to be later transformed into a rectangular shape should have its correlation coefficients contained in an "offset" $\mathbf{R}_t$ increased by a factor equal to $(3/\text{pi})^{1/2}$ (approximately 1.02333). The later transformation of one variable to a rectangular distribution will reduce the expected values of these "offset" coefficients to the desired values.

The multiplier to be applied to the offset coefficients in $\mathbf{R}_t$ when each of a pair of normal deviates are to be first transformed to achieve the desired expected corelations and later transformed to a rectangular distribution is less directly provided by Boldt. He provides an equation that can be used iteratively to obtain the required multiplier. Defining each correlation coefficient between two Gaussian distributed variables as $r_g$ and the correlation coefficient between the same two variables altered to assume a uniform (rectangular) distribution as $r_u$, we rewrite Boldt's formula (p. 2) as follows:

$$r_u = (6/\text{pi}) \text{ Arc Tan } [r_g/(4 - r_g^2]^{1/2} \quad . \tag{2}$$

We obtain our desired multiplier by discovering, by trial and error, the value of $r_g$ (the offset value) that will provide a value of $r_u$ equal to that we wish to have in $R_t$, our correlation matrix that defines the population we wish to sample in our model sampling experiment. The multiplier we use to produce the values in the "offset" matrix for those pairs of variables for which both variables are to be later transformed to rectangular distributions, is $r_g/r_u$.

Regardless of the convenience an off-the-shelf, or built-in generator may offer, researchers should avoid using any generator that does not permit the use of a designated seed permitting the repetition of the experiment. Also separate pseudorandom number generators should be used for each variable, each column of **X**. Our own program for creating random normal deviates uses a separate generator for each random number used to provide the vector of scores for each entity; these scores made up one row of **X**. For each synthetic normal deviate score we generate a uniformly distributed pseudorandom number, convert to an approximately Gaussian distributed number using a table look-up procedure, and then aggregate a number of these numbers to form the score for a specific variable in the score vector for an entity. We believe we are indulging in overkill to use so many independent components to constitute a score, but have found this program both affordable and reassuringly valid.

## E. MODEL SAMPLING RESEARCH DESIGN ISSUES

### 1. The Model Sampling Study

Model sampling studies are generally of two types: (1) evaluation of statistical methodology (e.g., the testing of robustness and distribution characteristics--Harris (1966) and Shields (1978), or (2) evaluation of utility of alternative research and operational strategies and procedures relating to the selection and assignment of personnel. We will focus on the latter type, in which synthetic scores provide the input into simulated personnel system models and the output provides a basis for determining the utility of alternative approaches.

Model sampling experiments are usually embedded in studies which, after the initial systems analysis and problem formulation stage, include the following five steps:

(1) identify and compute population values for the variables that provide the input into the simulation; populations of interest may comprise "youths," applicants, assignees, trainees, workers, candidates for promotion, etc.;

(2) generate synthetic scores for predictor and performance variables;

(3) simulate significant aspects of the personnel utilization process (i.e., selection, assignment, performance, performance evaluation, career decisions, reassignment, etc.);

(4) determine results (the output of the simulation), usually including the computation of MPP standard scores and the extent to which policy goals are met;

(5) analyze and interpret results; the output of each simulation becomes the unit of analysis for the testing of hypotheses and interpretation of results (including conversion of results into utility measures where appropriate).

Steps 2, 3, and 4 constitute the actual model sampling experiment. The experimental conditions contrasted for the testing of hypotheses may be reflected in characteristics of the entities generated in step 2 or in the decision processes of step 3. For example, alternative methods for selecting tests or test composites create different sets of variables which, in turn, require the generation of different sets of entities. A comparison of alternative selection/assignment algorithms might call for the use of the same entities for all experimental conditions; in such an experiment the experimental conditions may be distinguished by the use of different processes (i.e. separate simulations) in step 3 with the separate paths continued into step 4. Several model sampling designs are described in the appendices of Chapter 4.

Model sampling has both advantages and disadvantages when compared with a methodology that simulates a personnel utilization process using associated records of real persons obtained from a data bank. The data bank source of entities often seems more credible to managers who lack familiarity with model sampling; this approach provides scores comprising distributions that actually occurred in selecting from an applicant population, and thus the process for obtaining the scores can be understood readily without recourse to statistical theory. While the rejected applicants are not usually present in such data banks, the upper end of the test score distributions approximate that of applicants. On the negative side, the selected applicants have emerged as a result of both recruiting policies and selection practices that may or may not be continued into the future.

A major advantage of model sampling is that a youth population can be generated and the applicant population determined as the consequence of proposed recruiting policies; similarly, selection standards can be lowered to let in less qualified applicants or otherwise modified to make it compatible with recruiting strategies and/or requirements of the simulated system. Also, complex research designs using many independent samples can be utilized without making any one sample too small to provide stability of results.

Some of the detail provided by a data bank, while representative of the past, may have little chance of being replicated in the future in view of changing policies. Also the number of independent samples representing the desired population that can be drawn from a data bank frequently precludes the use of research designs that require many moderately large samples and some types of repeated measure designs. On the other hand, the availability of large data banks to determine empirical relationships can greatly improve the usefulness of the model sampling approach.

## 2. Entities for Model Sampling Experimentation

The researcher sometimes has the option of directly generating vectors of LSE scores or vectors of test scores that can be converted to LSEs; the entity is typically defined by its score vector. We will consider a type of model sampling experiment in which predicted performance scores (LSEs) are used to make all the decisions in the simulation of alternative selection/classification processes, and the MPP standard scores of those assigned to jobs are used as the result. In such an experiment the entities can consist of either predictor (test) scores or predicted performance scores.

Using the same notation as in section B above, $\mathbf{X}_n\mathbf{F}_t' = \mathbf{Y}$, and $\mathbf{Y}(\mathbf{R}_t^{-1})\mathbf{V}' = \mathbf{Z}$. Thus $\mathbf{Z} = \mathbf{X}_n\mathbf{F}_t'(\mathbf{R}_t^{-1})\mathbf{V}'$. Alternatively, $\mathbf{Z} = \mathbf{X}_n\mathbf{F}'$, where $\mathbf{F} = \mathbf{V}\mathbf{F}_t'(\mathbf{F}_t'\mathbf{F}_t)^{-1} = \mathbf{VB}(\mathbf{D}_b)^{-1/2}$. As noted in chapter 2, $\mathbf{F}$ can be defined in terms of the equation $\mathbf{C} = \mathbf{FF}'$, where $\mathbf{C}$ is the $m$ by $m$ matrix of covariances among the predicted performance scores of the $m$ jobs. In order to make a useful distinction in the present discussion we will define the $\mathbf{F}$ based on a factorization of $\mathbf{C}$ as $\mathbf{F}_c$ and note that $\mathbf{F}_c$ is an orthogonal rotation of the $\mathbf{F}$ defined as $\mathbf{VB}(\mathbf{D}_b)^{-1/2}$. We can now say that $\mathbf{Z} = \mathbf{X}_m\mathbf{F}_c'$ and that $E(\mathbf{X}_m\mathbf{F}_c') = E(\mathbf{X}_n\mathbf{VB}(\mathbf{D}_b)^{-1/2}$. When $m > n$, $\mathbf{F}_c$ will have $m - n$ null columns but the non-null columns of $\mathbf{F}_c$ will be within an orthogonal rotation of $\mathbf{F}$ defined as a factor extension of $\mathbf{F}_t$. When $n > m$, $\mathbf{Z}$ can be obtained more economically by generating $\mathbf{X}_m$ ($m$ random numbers per entity) and generating $\mathbf{Z}$ as equal to $\mathbf{X}_m\mathbf{F}_c'$.

The matrix $\mathbf{R}_t$ representing the population intercorrelations among n predictor variables may not be positive definite (have $n$ positive non-zero roots). Worse, since this matrix may have been computed on a sample that included incomplete data on some variables, or have been compiled from several sources including a few judicious estimates, $\mathbf{R}_t$ may not even be positive semidefinite (i.e., has at least one negative root proving that $\mathbf{R}_t$

could not have resulted from $\underline{Y}'\underline{Y}$ where the elements of $\underline{Y}$ are real numbers). In either instance it will usually be possible to select the $k$ largest factors having positive roots as a replacement for $\mathbf{F}_t$. In most cases this $n$ by $k$ factor matrix $\mathbf{F}_{tk}$ will adequately reproduce $\mathbf{R}_t$. The matrix reproduced by $\mathbf{F}_{tk}$ can, in most situations, be considered to be a better estimate of the universe values than is the original version of $\mathbf{R}_t$.

When an adequate reproduction of $\mathbf{R}_t$ is provided by $k$ factors, $k < n$, the synthetic test scores can be provided from pseudorandom normal deviates, that is, $\mathbf{Y} = \mathbf{X}_k\mathbf{F}_{tk}'$, where $\mathbf{Y}'\mathbf{Y}$ is equal to the redefined population $\mathbf{R}_t$, an intercorrelation matrix equal to $\mathbf{F}_{tk}\mathbf{F}_{tk}'$. Similarly, if $\mathbf{C}$ is adequately reproduced by the largest $k$ columns of $\mathbf{F}_a$, $\underline{\mathbf{Z}}$ is appropriately equated to $\mathbf{X}_k\mathbf{F}_{ak}'$.[21]

## 3. The Repeated Measure Design in Model Sampling

We will first consider a model sampling experiment in which the conditions being contrasted are all process related, that is, one in which none of the experimental contrasts requires the use of different entities to express different conditions. In such an experiment, each sample of entities can be used at each level of each treatment (factor) in the experimental design, and the full benefits can be realized from using a repeated measure design to reduce error variance. All the benefits that can accrue in the traditional psychological experiment from having each subject be his own control can be realized using entities instead of human subjects.

Experiments in which alternative methods of test selection are used in determining operational test batteries will usually require separate sets of variables to represent each selection method (experimental condition). Such an experiment requires the production of separate replication sets of entity samples reflecting each experimental condition. The model sampling procedure for creating one set of entity samples for each condition can use either: (1) a single vector of pseudorandom normal deviates separately transformed into each entity representing one experimental condition or, (2) separate vectors of pseudorandom normal deviates generated as the first step in producing each entity. The first procedure can provide highly correlated entities if the "best" transformation is used. The second procedure would provide completely independent entities. The use of correlated entities, as provided by the first method, will result in smaller differences being statistically significant.

[21] $\mathbf{F}_{tk}$ is a k factor PC solution of $\mathbf{R}_t$ and $\mathbf{F}_{ak}$ is a k factor PC solution of C.

When experimental conditions are expressed by the use of different variables in the simulated personnel procedures, all variables can usually be defined in terms of the same factor solution or in terms of factors that are within an orthogonal rotation of each other. Thus two test score vectors transformed from the same random normal deviate vector may be used as two separate entities in the simulation (to express two different experimental conditions), but these two entities behave statistically as if they were aspects of the same artificial person. A repeated measure design is then appropriate for the analysis of simulation results since this approach can reduce the error term to the same extent as the use of the same subject under two experimental conditions.

To illustrate the above example, we will consider two sets of $N$ entities, each consisting of one half of a set of ten synthetic test scores. $\mathbf{Y}_1$ is one $N$ by 5 matrix of synthetic test scores and $\mathbf{Y}_2$ is the second $N$ by 5 matrix of test scores. Both $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are transformed from the same $N$ by 5 $\mathbf{X}_n$ matrix (n equals 5). All n test scores are obtained from the equation, $\mathbf{Y} = \mathbf{X}_n F_t'$; each row of the $N$ by $n$ matrix $\mathbf{Y}$ represents an artificial individual and each column of $\mathbf{Y}$ represents a particular test. Each submatrix, both $\mathbf{Y}_1$ and $\mathbf{Y}_2$, consists of 5 columns of $\mathbf{Y}$ with the same row of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ representing a single individual. If the researcher desires independence between $\mathbf{Y}_1$ and $\mathbf{Y}_2$, he or she need only obtain $\mathbf{Y}_1$ and $\mathbf{Y}_2$ using two separate $\mathbf{X}_n$ matrices separately transformed into $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The implementation of such alternataive approaches is described in more detail in the appendices of Chapter 4.

In some sampling experiments the primary contrast may be between the results obtained from using factor scores corresponding to two separate factor solutions. Unless (or until) the factors making the smallest contributions are deleted, most alternative factor solutions will be within an orthogonal rotation of each other (obviously, oblique solutions are one important class of exceptions). Prior to deletion of these almost null (or complex) factors from the alternative orthogonal factor solutions of either $\mathbf{R}_t$ or $\mathbf{C}$ (using ones in the diagonal of $\mathbf{R}_t$ and the multiple correlation coefficients as the diagonal elements of $\mathbf{C}$) are orthogonal rotations of each other.

As noted in an earlier chapter, an orthogonal rotation of factors, factor scores, or test scores will not change the PCE resulting from their use in an assignment process. In other words, the same PCE attaches to all sets of scores that are within an orthogonal rotation of each other. Clearly, the use of different sets of factor scores obtained from transforming the same random vector is also analogous to using a subject as his own control, just as in the above example in which the entities consisted of test scores.

However, to the extent that the factors with small contributions are deleted after rotation to achieve a derived pc solution, as when $k$ factors are retained in a study of the PCE obtainable from factor scores using $\mathbf{F}_a$ compared to a set using $\mathbf{F}_d$, the entities based on $\mathbf{F}_a$ will have their correlation with the entities based on $\mathbf{F}_d$ reduced and the advantage of using a repeated measure design correspondingly reduced.

While two entities created as 1 by $k$ vectors of factor scores to represent $\mathbf{F}_a$ and $\mathbf{F}_d$ respectively are less correlated because the deletion of $n - k$ factors has removed different parts of the total space, the score vectors corresponding to both entities can be obtained from transforming the same 1 by $k$ vector of random numbers. The power of the statistical tests for determining the significance of MPP scores across cells can usually be increased by using a common $X_k$ for one replication in each cell of the results matrix, when the number of levels for each treatment are small. On the other hand, the stability of the estimate of the grand mean, as well as the means of treatments (factors) having a large number of levels, is increased by the use of independent random vectors to produce each entity. When the output of MPP standard scores is to be used to conduct a utility study, the stability of means is of paramount interest.

### 4. Flexibility in Modeling The Real World

Model sampling is a tool which can provide an investigator capability to control sources of bias and to make nontraditional assumptions that would not be feasible to apply if the investigator was restricted to the use of empirical data. Many of the standard assumptions used in traditional statistical experimental designs and statistical analyses are essential to the tractability of derivations and/or to conserve or maximize the information obtainable from scarce data while permitting the use of practical computing methods. Other models of the real world involving assumptions which conflict with those of the more traditional ones are equally attractive and could conceivably turn out to be more valid.

Compare the use of synthetic scores created by a model sampling process with the use of empirical scores drawn from an empirically created data bank to simulate, in both cases, an optimal classification system. In either case the assignment variables should be predicted performance (PP) estimates (i.e., LSEs) based on all the predictors. We call these variables "full least square (FLS) composites." After the assignment variables are used to optimally assign the entities to jobs, the benefits are measured in terms of mean predicted performance (MPP). FLS composites of the same form as those used for the assignment variables are used for evaluation--to compute the MPP of the assigned entities in each independent cross sample. The investigator should provide independent estimates of the weights for the variables comprising the FLS composites used for assignment and evaluation.

The research design summarized above also requires, for the conduct of the simulation and to compute MPP scores, the use of a sample, or samples, that is independent of both the analysis and evaluation samples. This is the traditional concept in which the validity of a "best" weighted composite is obtained in a different, independent, sample (i.e., the "cross" sample) as contrasted to the "back" sample on which the weights were computed. In summary, this research design requires: (1) two "back" samples, one (the analysis sample) on which to compute the weights for the assignment variables, and one (the evaluation sample) on which to compute the weights for the evaluation variables; and (2) one or more independent "cross" samples to be used for the conduct of the classification system simulation and computation of the MPP scores.

A simulation experiment using empirical scores requires dividing the total set of entities into separate analysis and evaluation samples while holding out enough entities for use as "cross" samples for the actual conduct of the simulation of the selection-classification process--unless the investigator has other prior, independent results from which to derive the required weights. A model sampling experiment has designated population parameters that are used to generate an analysis sample and to generate as many cross samples as desired. The investigator using model sampling to conduct his simulation does not need an evaluation sample, since the evaluation weights are appropriately computed using the designated population parameters. However, the investigator has the option of generating an evaluation sample if, for example, he or she wishes to replicate an empirical study which utilized such a sample.

The most commonly used model for the depiction of classification effects on MPP is one in which predicted performance scores are substituted for criterion (performance) scores. The use of this model does not require knowledge of the intercorrelations among the criterion variables while accurately depicting the relationships among predictors and between predictors and the criteria. This model appears appropriate for most selection-classification system simulations. However, alternative theories to be explored through model sampling experiments may stipulate relationships among the criterion variables, instead of among predictor variables. The joint predictor-criterion space may then be defined by extension of the criterion space to the predictor space rather than the more usual extension of the predictor space into the criterion space. Model sampling techniques for implementing such a criterion-based model are described in the Chapter 4 appendices.

The reliabilities of predictors, either individually or as sets, can be allowed to affect the characteristics of synthetic scores generated by model sampling. When predictions are considered as sets (i.e., as operational batteries) alternative concepts of reliability can be used to generate sets of parallel predictor forms. For example, using a true score plus error model of the total test score, the error component may, or may not, be correlated within a set of parallel forms. Either of these alternative models can be readily used as the basis of model sampling. Details for implementing alternative model sampling procedures for simulating parallel forms are provided in the Chapter 4 appendices.

## F. THE UNIVERSE: HOW TO DEFINE POPULATIONS IN TERMS OF COVARIANCE MATRICES

### 1. Unreliability of Criterion Measures

The first step in defining the population to be used as the basis for generating synthetic scores is to correct the covariances involving criterion variables for criterion unreliability. Such a correction is especially important when the criterion is comprised of several components that have different reliabilities and utility.

A correction for unreliability is easily accomplished as the traditional "correction for attenuation" when the component is the type of measure where reliability is a function of the number of raters or the length of a job competency or knowledge test. In such a measure the criterion used in the validation study is of arbitrary reliability and both the reason and basis for correcting are clear.

There is another kind of criterion measure for which corrections for unreliability are controversial. The concept of promotability can be defined as the ability of an individual to perform well at the next higher grade. This underlying capability can be thought of as a continuous variable which would correlate higher with predictors if it could be more reliably measured. If the concept is implemented as rated capability to perform at the higher grade, this is certainly true. If it is instead measured in terms of who is actually promoted, both the methodology for the correction and the justification for making such a correction come into question.

Thus, we would correct validities and criterion related covariances for unreliability as the first step in creating values of $\mathbf{R}_t$ and $\mathbf{V}$ or $\mathbf{C}$ to define the population on which a model sampling study is based. We prefer criterion measures which permit the computation of reliability measures but recognize that measures such as incarceration, promotion, or promotion rates, while philosophically not perfectly reliable measures of the underlying variable, do not lend themselves to a correction for attenuation.

### 2. Selection Effects

We do not believe it is conservative to use sample covariance values as representative of the population without applying corrections for selection effects. There are many ways that the relationships among and between predictors and the criteria can be distorted by the effects of restriction in range. The need to correct sample values for restriction in range should be particularly obvious when the results of many validity studies are being combined to provide relationships between experimental predictors and LSEs across all jobs. In these cases, each sample used foı validation will have different selection effects.

It is well known that when an operational battery is used to reject a significant number of applicants, the restriction in range effects will be more severe on these explicit selection variables than on those variables that are restricted only because they are positively correlated with the operational tests. In such a situation it is essential to use separate formulae for correcting the explicitly and incidentally selected variables. If no correction is used, or the same correction formula is used for all variables, there would be frequent replacement of operational tests with other tests that are not actually superior in an unrestricted population.

When operational tests have less of a role in selecting applicants than the interests and other self selection mechanisms exercised by the applicants, the designation of operational tests as explicit selectors will distort the validation results in favor of the operational tests to the disadvantage of new test content not already represented in the battery. Olson (1968) reports experimental data which show that applying such a model to the correction of the ACB provides too small a correction to experimental noncognitive tests (the measures believed to be most affected by self selection), making them less likely to be included in the operational battery.

Model sampling experiments can make use of data compiled from several sources. Also, the investigation of strategies for future research may call for using correlation matrices as an estimate of the population parameters that include biserial or tetrachoric coefficients, and/or coefficients based on patchwork samples with widely different Ns for different cells in the correlation matrix. Under such circumstances care must be taken not to correct biserial correlation coefficients as if they were product moment coefficients (biserial coefficients may be inflated by a restriction in range and should be inflated further by an improper correction), nor should an assembled matrix of coefficients be designated as the population matrix before being adjusted to assure that the matrix is positive semidefinite.

# APPENDIX 4A
# RESEARCH DESIGN ISSUES IN MODEL SAMPLING EXPERIMENTATION

## APPENDIX 4A.1: INTRODUCTION AND NOTATION

This appendix introduces and provides an initial, integrative, discussion of the methodological issues and techniques that are described in Appendices 4B and 4C. Material is placed in these appendices instead of the text because of their technical complexity and/or to provide additional detail not essential to the understanding of the chapter text or the evaluation of the Army classification system (Zeidner and Johnson, 1989b). We do not hesitate to express simulation techniques in the matrix algebra notation most useful for application and also make use of matrix notation to show the relationships between alternative approaches. Formal proofs are avoided.

The notation used in the appendices for this chapter follows:

$N$ = number of individuals or entities (synthetic individuals) used in a sample or group of entities considered together for selection and/or assignment purposes.

$n$ = number of predictor variables.

$m$ = number of jobs to which entities are to be assigned after selection and/or classification.

$\mathbf{X}_n$ = $N$ by $n$ matrix of normal deviates (synthetic scores); $E((\mathbf{X}_n{}'\mathbf{X}_n)/N)$ is equal to an $n$ by $n$ identity matrix ($\mathbf{I}_n$). $\mathbf{X}_m$ will be similarly used to denote an $N$ by $m$ matrix of normal deviate scores.

$\mathbf{Y}$ = $N$ by $n$ matrix of synthetic predictor scores (usually selection-classification tests) in standard score form; an underlined capital letter in bold print signifies that each score has been divided by the square square root of $N$.

$\mathbf{Z}_u$ = $N$ by $n$ matrix of criterion scores generated as synthetic normal deviates in standard score form.

$\mathbf{R}_t$ = $((\mathbf{Y}'\mathbf{Y})/N)$; an $n$ by $n$ matrix of correlation coefficients among the predictor variables. Using alternative notation, $\underline{\mathbf{Y}}'\underline{\mathbf{Y}} = \mathbf{R}_t$.

$\mathbf{Y}_J$ = An $N$ by $n$ matrix of predictor deviate scores in joint predictor-criterion space. These variables have the same correlations with the criterion scores as do the predictor scores making up the **Y** matrix but have reduced standard deviations and different intercorrelations.

$\mathbf{C}_J$ = $((\mathbf{Y}_j'\mathbf{Y}_j)/N)$; an $N$ by $N$ matrix of covariances among the predictor variables in the joint predictor- criterion space. This matrix has the same relationship to the predictor variables as $\mathbf{C}_p$ has to the criterion variables -- both relate to a set of variables existing empirically in a larger space (e.g., test space or criterion space) whose covariances are expressed in the joint predictor-criterion space.

$\mathbf{S}_J$ = An $n$ by $n$ diagonal matrix with the same diagonal elements as $\mathbf{C}_J$; these non-zero elements are the variances of the predictor variables in joint space.

$\mathbf{R}_J$ = $\mathbf{S}_J^{-1/2}(\mathbf{C}_J)\mathbf{S}_J^{-1/2}$; an $n$ by $n$ matrix of intercorrelations among the predictors in joint space.

**Z** = An $N$ by $m$ matrix of predicted performance (PP) deviate scores; each PP variable has a standard deviation (SD) equal to the multiple correlation of the specified set of predictor variables with the corresponding criterion variable.

$\mathbf{C}_p$ = $((\mathbf{Z}'\mathbf{Z})/N)$; an $m$ by $m$ matrix of covariances among the PP scores; the diagonal elements are multiple correlation coefficients.

$\mathbf{S}_p$ = An $m$ by $m$ diagonal matrix whose diagonal elements are the same as the diagonal elements of $\mathbf{C}_p$.

$\mathbf{R}_p$ = $\mathbf{S}_p^{-1/2}(\mathbf{C}_p)\mathbf{S}_p^{-1/2}$; an $m$ by $m$ matrix of correlation coefficients among the PP variables; the intercorrelations among the criterion variables in the joint predictor-criterion space (compare with $\mathbf{R}_J$).

$\mathbf{R}_u$ = An $m$ by $m$ matrix of intercorrelations among criterion variables.

**V** = an m by $n$ matrix of correlation coefficients between predictor and criterion scores; $V = Z_u'Y$; the same results are obtained if PP scores are substituted for criterion scores: $V = ((\mathbf{S}_p^{-1/2}\mathbf{Z}'\mathbf{Y})/N) = ((\mathbf{S}_p^{-1/2}(\mathbf{Z}'\mathbf{Y}_J)\mathbf{S}_J^{-1/2})/N)$; this is the validity matrix. **Q** = An N by k matrix of factor scores (expressed as standard scores).

## APPENDIX 4A.2: RESEARCH DESIGN ISSUES

Chapter 4 appendices focus on research designs directed at the evaluation of alternative selection-assignment policies that can be described in terms of actions taken on entity samples drawn from universes defined in terms of $\mathbf{R}_t$ and **V**. In the kind of model sampling experiments we envisage, investigators will usually make use of the supermatrix:

$$\begin{bmatrix} \mathbf{R}_t & \vdots & \mathbf{V}' \\ \cdots & \vdots & \cdots \\ \mathbf{V} & \vdots & \mathbf{R}_p \end{bmatrix}$$

or, much more rarely,

$$\begin{bmatrix} \mathbf{R}_t & \vdots & \mathbf{V}' \\ \cdots & \vdots & \cdots \\ \mathbf{V} & \vdots & \mathbf{R}_u \end{bmatrix} .$$

In studies where the benefits side of utility is calculated from mean predicted performance (MPP) resulting from the implementation of each policy, there is no need to know the covariances among actual criterion scores. However, we briefly consider an alternative model in which knowledge of $\mathbf{R}_u$ is available and useful.

Sets of predictor and criterion scores generated from universe values of $\mathbf{R}_t$ and $\mathbf{V}$ can be assumed to be: (1) universe values, (2) analysis ("back") samples used to select variables and compute weights to apply to decision variables of simulated systems, (3) evaluation samples used to compute weights for evaluation variables to apply to cross sample scores, or (4) the independent ("cross") samples that provide the predictor scores to which the various weights or other parameter values are applied. Some psychometric studies may require only one or more samples that can be assumed to be best estimates of the universe. However, the comparison of alternative personnel policies will usually require more complicated research designs to prevent the biasing of results by correlational error (as betweem LSEs of predicted performance used as assignment versus evaluation variables).

A basic research design calls for using a value for $\mathbf{R}_t$ and $\mathbf{V}$ to generate a set of $\mathbf{Y}$ and $\mathbf{Z}$ matrices from which the analysis sample values can be computed. Weights to be applied to selected cross sample predictor scores to provide LSEs of PP scores to be used as decision variables (e.g., for selection and assignment) are contained in an $n$ by $m$ matrix, $\mathbf{W}_a$. Similarly, the weights to be applied to the same cross sample predictor scores to provide FLS estimates of PP scores to be used as evaluation variables are contained in an $n$ *by* $m$ matrix, $\mathbf{W}_e$. We can compute these matrices as: $\mathbf{W}_a = \mathbf{R}_{ta}^{-1}\ \mathbf{V}_a'$, and $\mathbf{W}_e = \mathbf{R}_{te}^{-1}$ $\mathbf{V}_e'$. In each cross sample the policy decisions can be made on simulated decisions made on the basis of $\mathbf{YW}_a$, and the results measured in terms of predicted performance scores computed from "universe" or evaluation sample values to compute $\mathbf{W}_e$ (to apply to cross sample values of $\mathbf{Y}$ to yield $\mathbf{Z}_e = \mathbf{Y}\mathbf{W}_e$.

Some of the research questions that could be answered using the above model sampling research design include:

(1) Which technique is best for selecting predictor variables (e.g., for selecting operational selection-classification batteries, for selecting sets of test composites and corresponding job families)?

(2) Which personnel system characteristics are most efficient (e.g., equal versus disparate variances for aptitude areas, quality distribution across job families, sets of prerequisite or cutting scores)?

(3) Which selection-classification algorithms (processes) have the most positive effect on MPP (e.g., one versus two stage selection-classification systems, use of MDS and/or person-by-person assignment algorithms)?

The "cross" samples of entities can be used in repeated measure analysis of variance designs to obtain maximum sensitivity to policy effects. Or conversely, since the model sampling approach permits the generation of an unlimited number of completely independent samples, the investigator can afford to use completely distinct, independent, samples across conditions whenever he so prefers.

Where LP algorithms are used to optimally assign entities to jobs, it will usually be practical to make use of several replications of relatively small samples, possibly an $N$ of between 200 to 300 when $m < 15$. There is considerable evidence that twenty samples of $N = 200$ provides alsmost as much stability as a single sample of 4,000, but the cost of optimally assigning 4,000 entities to 9 job families in a single solution is many times over that of making 20 such solutions in samples of 200 each.

The identification of values for $\mathbf{R}_t$ and $\mathbf{V}$ to define a desired universe will usually require correcting empirically obtained correlation coefficients with restriction in range (correction for selection effects) formulae. Depending on the study, the described universe may consist of the youth, applicant trainee, on-the-job, second term, or career populations. The empirical data will usually have been collected on a sample drawn from the on-the-job population.

Some of the above topics will be explored in more detail in the remaining appendices of this chapter. The generation of predictor, factor, and PP scores will be described in Appendix 4B. The independence of policy decision and evaluation variables will be explained further in Appendix 4C.

# APPENDIX 4B
# GENERATING SYNTHETIC SCORES

## APPENDIX 4B.1: GENERATING PREDICTOR OR PP SCORES FROM $X_n$ OR $X_m$

The objective of a model sampling experiment might be to determine which of several sets of predictors will provide the maximum amount of MPP under the condition of optimal assignment. If the investigator makes the assumption that he knows the parameters of the universe (defined by $\mathbf{R}_t$ and $\mathbf{V}$), the model sampling experiment becomes a substitute for being able to analytically solve a set of definite integrals of a multivariate normal density function. In this case these functions would have $\mathbf{C}_p$ as a parameter. The experiment would thus be a means of solving an otherwise implacable mathematical problem.

Since the personnel operations being simulated include optimal assignment, an objective function (i.e., MPP) is maximized in each such replication. In contrast to the more general research design described in Appendix 4A.2, this objective function produced as a product of the assignment algorithm (the allocation sum or MPP standard score) is also the estimate of MPP used as the evaluation variable used as the benefits component for computing utility.

In many, if not most, such designs each experimental condition can be represented by a particular $m$ by $m$ matrix $\mathbf{C}_p$. Assuming the more difficult circumstances of $m < n$, an economy of effort can be achieved by the generation of the cross sample $\mathbf{Z}$ matrices used for both assignment and evaluation from $\mathbf{X}_m$ rather than from $\mathbf{X}_n$. Using the principles and relationships described in the Chapter 2 appendices, we can define $\mathbf{Z}$ in terms of the following sequences:

**First Sequence**

$\mathbf{Z} = \mathbf{YW}$; $\mathbf{Y} = \mathbf{X}_n \mathbf{R}_t^{1/2}$; $\mathbf{W} = \mathbf{R}_t^{-1} \mathbf{V}'$; thus

$$\mathbf{Z} = \mathbf{X}_n \mathbf{R}_t^{1/2} \mathbf{R}_t^{-1} \mathbf{V}' = \mathbf{X}_n \mathbf{R}_t^{-1/2} \mathbf{V}', \text{ and} \tag{1a}$$

where $\mathbf{A}_t \mathbf{D}_t \mathbf{A}_t' = \mathbf{R}_t$, and $\mathbf{A}_t \mathbf{A}_t' = \mathbf{A}_t' \mathbf{A}_t = \mathbf{I}_n$,

$$\mathbf{Z} = \mathbf{X}_n (\mathbf{V} \mathbf{A}_t \mathbf{D}_t^{-1/2})' = \mathbf{X}_n \mathbf{D}_t^{-1/2} \mathbf{A}_t' \mathbf{V}'. \tag{1b}$$

**Second Sequence**

$\mathbf{Z} = \mathbf{X}_m \, \mathbf{C}_p^{1/2}$, and considering that: (2a)

$\mathbf{C}_p = \mathbf{F}_v\mathbf{F}_v' = \mathbf{F}_p\mathbf{F}_p'$, where $\mathbf{F}_v$ is an $m$ by $n$ factor extension of $\mathbf{F}_t$ into the joint space, and $\mathbf{F}_p$ is an $m$ by $m$ PC factor solution of $\mathbf{C}_p$, and

$\mathbf{F}_v \, \mathbf{A}_p = \mathbf{F}_p$, where $\mathbf{A}_p'(\mathbf{F}_v'\mathbf{F}_v)\mathbf{A}_p = \mathbf{D}_p$, then,

$$\mathbf{Z} = \mathbf{X}_m \, (\mathbf{F}_v \, \mathbf{A}_p)' = \mathbf{X}_m \, \mathbf{A}_p'\mathbf{D}_t^{-1/2} \, \mathbf{A}_t'\mathbf{V}'. \qquad (2b)$$

It is readily seen that the **Z** matrices resulting from the above two sequences both have the relationship that $E((\mathbf{Z}'\mathbf{Z})/N) = \mathbf{C}_p$. It is clear that either sequence can be used for a model sampling experiment in which samples are drawn from the universe to make mathematical computations. When parameters from independent analysis and evaluation samples (e.g., $\mathbf{W}_a$ and $\mathbf{W}_e$) are to be used, it is necessary to use either the first sequence or the third sequence to be described below.

When $n$ is considerably larger than $m$, there can be practical value in using $\mathbf{X}_m$ instead of $\mathbf{X}_n$ to first create predictor scores, apply $\mathbf{W}_a$ to cross sample predictor scores as a means of creating LSEs for use as assignment variables, and then applying $\mathbf{W}_e$ to the cross sample predictor scores for use as evaluation variables. We will demonstrate the feasibility of using $\mathbf{X}_m$ to generate predictor scores in the joint predictor-criterion space for use in creating separate **Z** matrices for use in: (1) the simulation decision process, and (2) the evaluation process.

## APPENDIX 4B.2: GENERATING SCORE MATRICES FOR PREDICTOR VARIABLES IN THE JOINT PREDICTOR-CRITERION SPACE

The $N$ by $n$ matrix of predictor scores in the joint space is denoted as $\mathbf{Y}_J$. By definition, $((\mathbf{Y}_J'\mathbf{Y}_J)/N)$ equals $\mathbf{C}_J$, the covariances of the predictors in the joint predictor space. Defining $\mathbf{S}_J$ as a diagonal matrix whose non-zero elements are the diagonal elements of $\mathbf{C}_J$, we can define $\mathbf{R}_J$ as $S_J^{-1/2}(C_J)S_J^{-1/2}$, and $\mathbf{Y}_{J1} = \mathbf{Y}_J \, S_J^{-1/2}$; thus, $(\mathbf{Y}_{J1}' \mathbf{Y}_{J1})/N = \mathbf{R}_J$. Also, $\mathbf{V}' = (S_J^{-1/2}(\mathbf{Y}_J'\mathbf{Z})S_p^{-1/2})/N)$, where $S_p$ is the diagonal matrix with the same diagonal elements as $\mathbf{C}_p$.$\mathbf{Y}_{J1} \, \mathbf{W}_a = \mathbf{Z}_a$, providing the PP scores used as assignment variables and the equation $\mathbf{Y}_{J1} \, \mathbf{W}_e = \mathbf{Z}_e$, provides the PP scores used to compute the MPP standard scores used for evaluation.

To make use of development logic provided in the appendices of Chapter 2, we first substitute $\mathbf{F}_p'$ for $\mathbf{C}_p^{1/2}$ in formula (2a) as the means of transforming $\mathbf{X}_m$ into **Z**, thus providing formula (2b). The factor extension of $\mathbf{F}_p$ into the predictor space can be expressed as:

$$\begin{bmatrix} \mathbf{R}_t \\ \cdots \\ \mathbf{V} \end{bmatrix} \mathbf{T} = \begin{bmatrix} \mathbf{F}_{tp} \\ \cdots \\ \mathbf{F}_p \end{bmatrix} . \qquad (3)$$

The $n$ by $m$ matix $\mathbf{T}$ can be developed sequentially as follows:

$$\begin{bmatrix} \mathbf{R}_t \\ \cdots \\ \mathbf{V} \end{bmatrix} \mathbf{T}_1 = \begin{bmatrix} \mathbf{F}_t \\ \cdots \\ \mathbf{F}_v \end{bmatrix} ; \quad \begin{bmatrix} \mathbf{F}_t \\ \cdots \\ \mathbf{F}_v \end{bmatrix} \mathbf{T}_2 = \begin{bmatrix} \mathbf{F}_{tp} \\ \cdots \\ \mathbf{F}_p \end{bmatrix} ;$$

$\mathbf{T} = \mathbf{T}_1\mathbf{T}_2$, an $n$ by $n$ by an $n$ by $m$ matrix yielding an $n$ by $m$ matrix as a product.

In this example $\mathbf{F}_t$ is the PC factor solution of $\mathbf{R}_t$; $\mathbf{F}_t = \mathbf{R}_t\, \mathbf{A}_t\, \mathbf{D}_t^{-1/2}$, where $\mathbf{A}_t'\mathbf{R}_t\mathbf{A}_t' = \mathbf{D}_t$, $\mathbf{A}_t'\mathbf{A}_t = \mathbf{A}_t\mathbf{A}_t' = \mathbf{I}_n$, and $\mathbf{D}_t$ is the $n$ by $n$ diagonal matrix of eigen values. $\mathbf{T}_1$ is equal to $\mathbf{A}_t\mathbf{D}_t^{-1/2}$, as explained in the Chapter 3 appendices. We define $\mathbf{T}_2$ to be $\mathbf{A}_p$, where $\mathbf{A}_p$ is the eigen vector matrix and $\mathbf{D}_p$ is the eigen value (diagonal) matrix found in the unique equality, $\mathbf{A}_p'(\mathbf{F}_v'\mathbf{F}_v)\mathbf{A}_p = \mathbf{D}_p$, where $\mathbf{A}_p'\mathbf{A}_p = \mathbf{I}_m$ and $\mathbf{A}_p\mathbf{A}_p'$ definitely does not equal $\mathbf{I}_n$ ($\mathbf{A}_p$ is an $n$ by $m$ orthonormal matrix); $\mathbf{F}_p = \mathbf{F}_v\mathbf{A}_p$. We see that $\mathbf{F}_v$ is an $m$ by $n$ extended factor solution, in the sense of Dwyer, and $(\mathbf{F}_v'\mathbf{F}_v)$ is an $n$ by $n$ matrix with rank $m$ (i.e., can be completely described by $m$ factors).

As noted above, $\mathbf{T}$ is equal to $\mathbf{T}_1\mathbf{T}_2$ or $\mathbf{A}_t\, \mathbf{D}_t^{-1/2}\mathbf{A}_p$, an $n$ by $m$ matrix. $\mathbf{F}_{tp}$, obtainable as $\mathbf{R}_t\, \mathbf{T}$, is an $n$ by $m$ factor solution such that $\mathbf{X}_m\, \mathbf{F}_{tp}'$ yields $\mathbf{Y}_J$, an $N$ by $n$ matrix of predictor deviate scores in the joint predictor-criterion space. The scores in $\mathbf{Y}_J$ have many of the critical characteristics of the predictor scores in test space, including the relationship $((\mathbf{S}_J^{-1/2}\mathbf{Z}'\mathbf{Y}_J\mathbf{S}_J^{-1/2})/N) = \mathbf{V}$, that is, these predictor scores in joint space have the same correlations with the criterion scores as do the predictor scores in test space. However, the relationship of $\mathbf{Y}_J$ to $\mathbf{Y}$ is a complex one and needs to be explored systematically. We describe how these two sets of predictor scores are different and where they can be expected to provide the same results.

In considering the effect of using $\mathbf{Y}_J$ instead of $\mathbf{Y}$ we begin by defining $\mathbf{V}_J$ as the $m$ by $n$ matrix of correlation coefficients between PP scores and predictor scores when we substitute $\mathbf{Y}_J$ for $\mathbf{Y}$; we then define $\mathbf{R}_j$ as in formula (5):

$$\mathbf{V_J} = \mathbf{F_pF_{tp}}'; \mathbf{F_p} = \mathbf{V}\,\mathbf{R_t}^{-1}\,\mathbf{A_p}; \mathbf{F_{tp}} = \mathbf{R_t}^{1/2}\,\mathbf{A_p} \;;$$

$$\mathbf{V_J} = \mathbf{V}\,\mathbf{R_t}^{-1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{1/2} \quad . \tag{4}$$

$$\mathbf{R_J} = \mathbf{F_{tp}}\,\mathbf{F_{tp}}' = \mathbf{R_t}^{1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{1/2} \quad . \tag{5}$$

Before proceeding further, we need to consider some of the special properties of the matrix $(\mathbf{A_pA_p}')$. If the product of a matrix with itself equals that matrix, as in the equality, $\mathbf{M}^2 = \mathbf{M}$, then $\mathbf{M}$ is an idempotent matrix and it is easily verified that the generalized inverse of $\mathbf{M}$ (i.e., $\mathbf{M}^*$) is equal to $\mathbf{M}$, since $\mathbf{M} = \mathbf{M}\,\mathbf{M}^*\,\mathbf{M}$ when $\mathrm{M} = \mathrm{M}^*$. While $(\mathbf{A_pA_p}')$ is not the identity matrix, as is $(\mathbf{A_p}'\mathbf{A_p})$, the matrix $(\mathbf{A_pA_p}')$ is idempotent--almost, but not quite, as useful a property.

Since $\mathbf{R_J}$ is of order $n$ with rank $m$ and $m < n$, this matrix does not have an ordinary inverse; we do not have recourse to the matrix $(\mathbf{R_J})^{-1}$. However, the generalized inverse, $(\mathbf{R_J})^*$, does exist and looking at our formula for $\mathbf{R_J}$ we see immediately that we can write this generalized inverse as $\mathbf{R_t}^{-1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{-1/2}$. We noted above that $(\mathbf{A_pA_p}')^*$ equals $(\mathbf{A_pA_p}')$, and thus:

$$\mathbf{R_J}^* = \mathbf{R_t}^{-1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{-1/2} \quad . \tag{6}$$

The product of $\mathbf{R_J}$ and $\mathbf{R_J}^*$,i.e., $\mathbf{R_J}\,\mathbf{R_J}^*$ is seen to be equal to $\mathbf{R_t}^{+1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{-1/2}$. Returning to equation (4) we note that this expression for $\mathbf{V_J}$ can be written as follows:

$$\mathbf{V_J} = \mathbf{V}\,(\mathbf{R_J})^*\,\mathbf{R_J} \text{ ,and } \mathbf{V_J}' = \mathbf{R_J}\,(\mathbf{R_J})^*\,\mathbf{V}' \quad .$$

We can now write: $\mathbf{R_J}^*\,\mathbf{V}' = \mathbf{R_J}^*\,(\mathbf{R_JR_J}^*\,\mathbf{V}') = \mathbf{R_J}^*\,\mathbf{V}'$, since by the definition of $\mathbf{R_J}^*$, $\mathbf{R_J}^* = \mathbf{R_J}^*\,\mathbf{R_J}\,\mathbf{R_J}^*$ (as well as $\mathbf{R_J} = \mathbf{R_J}\,\mathbf{R_J}^*\,\mathbf{R_J}$ ).

We can also compare the two expected covariance matrices produced by changing the regression weight matrix, $\mathbf{W}$, in the expression $\mathrm{E}((\mathbf{Z}'\mathbf{Y_{J1}}\,\mathbf{W})/N)$. In the following two developments we compare the expected PP covariance matrices for $\mathbf{W} = (\mathbf{R_t})^{-1}\mathbf{V}'$ with $\mathbf{W} = (\mathbf{R_J})^{-1}\mathbf{V}'$.

Firstly, when $\mathbf{W} = \mathbf{Rt}^{-1}\mathbf{V}'$ we have:

$$\begin{aligned} \mathrm{E}(\mathbf{Z}'\mathbf{Y_{J1}}\,\mathbf{R_t}^{-1}\,\mathbf{V}') &= \mathbf{V}\,\mathbf{R_t}^{-1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{+1/2}\,\mathbf{R_t}^{-1}\,\mathbf{V}' \\ &= \mathbf{V}\,\mathbf{R_t}^{-1/2}\,(\mathbf{A_pA_p}')\,\mathbf{R_t}^{-1/2} \\ &= \mathbf{V}\,\mathbf{R_J}^*\,\mathbf{V}' \quad . \end{aligned}$$

Secondly, when $\mathbf{W} = \mathbf{R_J}^*\,\mathbf{V}'$ we have:

$$E(Z'Y_{J1}\,R_J^*\,V') = V\,R_t^{-1/2}\,(A_pA_p')\,R_t^{+1/2}\,R_J^*\,V' \;;$$

$$R_J^* = R_t^{-1/2}\,(A_pA_p')\,R_t^{-1/2} \;.$$

$$E(Z'Y_{J1}\,R_J^*\,V') = V\,(R_t^{-1/2}\,(A_pA_p')\,R_t^{-1/2})V'$$

$$= V\,(R_J^*)V' \;.$$

Our third sequence, as provided below, can be used in research designs where: (1) one set of estimates of PP scores or other decision variables are computed for simulation of personnel system or policy decisions; and (2) a second set of independent estimates are to be used in the evaluation of the simulation effects resulting from the simulation.

**Third Sequence:**

$$Y_J = X_m\,(T'R_t);\; T'R_t = A_p'\,D_t^{-1/2}\,A_t' \;. \qquad (7)$$

$$Z = Y_J\,(S_J)^{-1/2}\,W_J : W_J = (R_J)^{-1}\,V' \;. \qquad (8)$$

The $Y_J$ matrices generated for "cross" samples by formula (7) are appropriately obtained using values for $T'R_t$ computed from the universe values of $R_t$ and $V$, while the weights to be applied to these "test" scores should be computed in an independent analysis sample. The $Z$ matrices resulting from formula 8 are computed for each cross sample using either the universe analysis or evaluation sample values of $R_t$ and $V$ as appropriate for the research design (see Appendix 4C).

## APPENDIX 4B.3: GENERATING FACTOR SCORE MATRICES COMMENCING WITH EITHER $Y_J$ OR Y

The third sequence can be readily extended to generate an $N$ by $k$ matrix of rotated factor scores. The factor solution of $C_p$, $F_v$ can be rotated to alternative solutions (e.g., to simple structure), to show the correlation of each of the m PP variables with the $k$ rotated factors of $F_{vr}$. The transformation matrix relating $F_p$ to $F_{pr}$ is $T_r$, that is, $F_{vr} = F_v\,T_r$. The correlations of the predictor variables with these rotated factors is denoted as $F_{tr}$, where $F_v\,T_r = F_{tr}$.

The following formulae are repeated here for ready reference during the derivations and demonstrations of this section.

$$F_t = R^{1/2} \quad ; \quad F_{tp} = F_t\,A_p.$$

$$F_v = V\,R_t^{-1/2} \quad ; \quad F_p = F_v\,A_p.$$

$\mathbf{Y}_J = \mathbf{X}_m (\mathbf{F}_{tp})'$ ; $\mathbf{Y}_{J1} = \mathbf{Y}_J \mathbf{S}_J^{-1/2}$ ; $[(\mathbf{Y}_{J1})' \mathbf{Y}_{J1}]/N = \mathbf{R}_J$.

The desired matrix of factor scores, $\mathbf{Q}_r$, can be generated either from $\mathbf{X}_n$ or from $\mathbf{X}_m$ as follows:

**Generating factor scores from $\mathbf{X}_n$:**

$\mathbf{Q}_r = \mathbf{Y}\ \mathbf{R}_t^{-1}\ \mathbf{T}_r$, or in terms of $\mathbf{X}_n$,

$\mathbf{Q}_r = \mathbf{X}_n\ \mathbf{F}_t'\mathbf{R}_t^{-1}\mathbf{T}_r$, which in turn reduces to,

$\mathbf{Q}_r = \mathbf{X}_n\ \mathbf{R}_t^{-1/2}\ \mathbf{T}_r$; when factors are to be defined in terms of predictor variables in "total test space"; (9a)

**Generating Factor Scores From $\mathbf{X}_m$:**

$\mathbf{Q}_r = \mathbf{Y}_{J1}\ \mathbf{R}_J^{*}\ \mathbf{R}_t^{1/2}\mathbf{T}_r$, or in terms of $\mathbf{X}_m$,

$\mathbf{Q}_r = \mathbf{X}_m\ \mathbf{F}_{tp}'\mathbf{R}_J^{*}\ \mathbf{R}_t^{1/2}\mathbf{T}_r$, which reduces to,

$\mathbf{Q}_r = \mathbf{X}_m\ \mathbf{A}_p'\ \mathbf{T}_r$, when factors are to be defined in terms of predictor variables in "joint predictor-criterion space". (9b)

When $\mathbf{T}_r = \mathbf{A}_p$, $\mathbf{Q}_r = \mathbf{X}_m$. It is readily verified that $(\mathbf{Q}_r'\mathbf{Q}_r)/N = \mathbf{I}_m$, $(\mathbf{Z}'\mathbf{Q}_r)/N = \mathbf{F}_p$, and $(\mathbf{Y}_J'\mathbf{Q}_r)/N = \mathbf{F}_{tp}$, when $\mathbf{Q}_r$ is equal to $\mathbf{X}_m$. We next investigate the general case, when $\mathbf{T}_r$ does not equal $\mathbf{A}_p$, and demonstrate that all three of the desired properties hold in the general case.

An N by k matrix of factor scores, $\mathbf{Q}$, should have (or closely approximate when k < m), the following properties:

$E[(\mathbf{Q}_r'\mathbf{Q}_r)/N] = \mathbf{I}_k$ ,or equal to $\mathbf{R}_q$ if rotation is to oblique factors (10)

$E[(\mathbf{Z}'\mathbf{Q}_r)/N] = \mathbf{F}_{tr} = \mathbf{V}\ \mathbf{R}_t^{-1/2}\ \mathbf{T}_r$, where $\mathbf{F}_{vr} = \mathbf{F}_v\mathbf{T}_r$ (11)

and $\mathbf{F}_{tr} = \mathbf{F}_t\,\mathbf{T}_r$.

$E[(\mathbf{Y}'\mathbf{Q}_r)/N] = \mathbf{F}_{tr} = \mathbf{R}_t^{1/2}\ \mathbf{T}_r$ (12)

We suggest defining the factor scores in terms of either $\mathbf{Y}$ or $\mathbf{Y}_J$; a particular $\mathbf{Q}_r$ can be either equal to $(\mathbf{Y}\ \mathbf{R}_t^{-1}\ \mathbf{F}_{tr})$ or to $(\mathbf{Y}_J\ \mathbf{R}_J^{*}\ \mathbf{F}_{tr})$ with $\mathbf{F}_{tr}$ equal to $\mathbf{F}_t\mathbf{T}_r$. Since the credibility of using $\mathbf{Y}$ is in little doubt, where $\mathbf{Y} = \mathbf{X}_n\ \mathbf{F}_t'$, we demonstrate the adequacy of a $\mathbf{Q}_r$ based on $\mathbf{Y}_J$, a parallel demonstration of the adequacy of a $\mathbf{Q}_r$ based on $\mathbf{Y}$ would follow similar lines but would be much easier to produce, since the regular inverse is available for use in simplifying algebraic expressions.

Commencing with $\mathbf{Q}_r = \mathbf{Y}_J \mathbf{R}_J^* \mathbf{F}_{tr}$, we can readily first expand and then simplify the expression for $\mathbf{Q}_r'\mathbf{Q}_r/N$ as follows:

$[(\mathbf{Q}_r'\mathbf{Q}_r)/N] = \mathbf{F}_{tp}' \mathbf{R}_J^* \mathbf{R}_J \mathbf{R}_J^* \mathbf{F}_{tp}$

$= \mathbf{F}_{tp}' \mathbf{R}_J^* \mathbf{F}_{tp}$

$= \mathbf{A}_p'\mathbf{R}_t^{1/2}(\mathbf{R}_t^{-1/2} (\mathbf{A}_p\mathbf{A}_p') \mathbf{R}_t^{-1/2})\mathbf{R}_t^{1/2} \mathbf{A}_p$, first collapsing $\mathbf{R}_t^{1/2} \mathbf{R}_t^{-1/2}$ into $\mathbf{I}_n$, and then $\mathbf{A}_p'\mathbf{A}_p$ into $\mathbf{I}_n$, we simplify this total expression to $\mathbf{I}_n$.

$[(\mathbf{Q}_p\mathbf{Q}_p)/N] = \mathbf{I}_n$, the first of our three desired properties.

Noting that $\mathbf{Z} = \mathbf{Y}_J\mathbf{R}_J^* \mathbf{V}'$, $\mathbf{Q}_r = \mathbf{Y}_J \mathbf{R}_J^* \mathbf{F}_{tr}$, and $\mathbf{F}_{tp} = \mathbf{R}^{1/2} \mathbf{A}_p$, we now expand and then simplify the expression for $\mathbf{Z}'\mathbf{Q}_r/N$ as follows:

$[(\mathbf{Z}'\mathbf{Q}_r)/N] \quad = (\mathbf{V}\mathbf{R}_{J1}^* (\mathbf{Y}_{J1}'\mathbf{Y}_{J1}) \mathbf{R}_{J1}^* \mathbf{R}_t^{1/2} \mathbf{T}_r)/N$ ;

$= \mathbf{V} (\mathbf{R}_{J1}^* \mathbf{R}_{J1} \mathbf{R}_{J1}^*) \mathbf{R}_t^{1/2} \mathbf{T}_r$;

$= \mathbf{V} \mathbf{R}_t^{-1/2} \mathbf{T}_r = \mathbf{F}_v \mathbf{T}_r = \mathbf{F}_r$.

$[(\mathbf{Z}_1'\mathbf{Q}_r)/N] \quad = \mathbf{F}_{vr}$, and the second of the three desired properties in a matrix of factor scores is shown to be present.

The third desired property is shown in the following sequence to be also present in a $\mathbf{Q}_r$ generated fron $\mathbf{Y}_J$:

$[(\mathbf{Y}_{J1}'\mathbf{Q}r)/N] = (\mathbf{Y}_{J1}'\mathbf{Y}_{J1} \mathbf{R}_{J1}^* \mathbf{F}_{tr})/N$

$= (\mathbf{R}_{J1} \mathbf{R}_{J1}^*) \mathbf{F}_{tr} = (\mathbf{R}_{J1} \mathbf{R}_{J1}^*) \mathbf{R}_t^{1/2} \mathbf{T}_r$

$= (\mathbf{R}_t^{+1/2} (\mathbf{A}_p\mathbf{A}_p') \mathbf{R}_t^{-1/2})\mathbf{R}_t^{1/2} \mathbf{T}_r$

$[(\mathbf{Y}_{J1}'\mathbf{Q}_p)/N] = \mathbf{R}_t^{1/2} \mathbf{T}_r = \mathbf{F}_{tr}$.

All three of the desired properties are present for a $\mathbf{Q}_r$ defined as being equal to $\mathbf{Y}_J \mathbf{R}_{J1}^* \mathbf{F}_{tr}$. Thus these desired properites of a $\mathbf{Q}_r$ matrix are possessed by any orthogonal or oblique rotation applied to either $\mathbf{F}_t$, $\mathbf{F}_{tp}$ or directly to $\mathbf{Q}_p$ or $\mathbf{Q}_r$.

We see that factor scores based on either $\mathbf{Y}$ or $\mathbf{Y}_J$ can possess the desired properties of factor scores. The generation of either $\mathbf{Q}$ matrix is quite simple, since $\mathbf{Y} = \mathbf{X}_n \mathbf{F}_t'$ and $\mathbf{Y}_J = \mathbf{X}_m \mathbf{F}_{tp}'$, $\mathbf{Q}_r$ can be generated by either formula (9a), based on $\mathbf{Y}$ or on formula (9b), based on $\mathbf{Y}_J$. When $n > m$, there is a small economy effected by using $\mathbf{X}_m$ instead of $\mathbf{X}_n$. However, the primary advantage derived from being able to generate factor or other predictor scores, as well as PP scores and/or criterion scores, from the same $\mathbf{X}_m$ matrix is that some simulations may require the direct generation of criterion scores. A matrix of criterion scores, $\mathbf{Z_u}$, as contrasted to a matrix of PP scores, $\mathbf{Z}$, cannot be obtained as a function of the predictor scores--instead, a $\mathbf{Z_u}$ matrix must be generated from $\mathbf{X_m}$ and it is essential to generate the predictor and criterion scores from the same $\mathbf{X_m}$ or $\mathbf{X_u}$ (in this case it must be from $\mathbf{X_m}$).

## APPENDIX 4B.4: AN APPROACH THAT USES THE INTERCORRELATIONS AMONG THE CRITERION VARIABLES

The purpose of this section is to compare the meaning and usefulness of two alternative sets of variables--measures of predicted performance as compared to actual performance. While actual performance measures on several jobs for the same individual would be very rarely obtainable, hypothetical criterion universes might be credibly formulated for use in simulations. The use of either predicted performance or actual performance considered here emphasize the joint predictor-criterion space. The two alternatives being considered are: (1) the identification of factors in predictor space that are then extended into the criterion space, or (2) the identification of factors in criterion space that are then extended into the predictor space.

We believe the most useful approach to the conduct of model sampling research on the utility of selection and classification relates to the first of the two factor solutions shown below. However, the second of these two, an alternative model, can be used when the investigator is able to stipulate the correlations among the criterion variables and wishes to use this $m$ by $m$ matrix of correlations among these variables, the matrix $\mathbf{R_u}$, as the basis for computing the utility of personnel policies.

**Primary Model:**

$$\mathbf{F_a F_a'} = \left[\begin{array}{c:c} \mathbf{R_t} & \mathbf{V'} \\ \hdashline \mathbf{V} & \mathbf{C_p} \end{array}\right], \qquad \mathbf{F_a} = \begin{bmatrix} \mathbf{F_t} \\ \cdots \\ \mathbf{F_v} \end{bmatrix}.$$

**Alternative Model:**

$$\mathbf{F_b F_b'} = , \left[\begin{array}{c:c} \mathbf{R_u} & \mathbf{V} \\ \hdashline \mathbf{V'} & \mathbf{C_J} \end{array}\right] \qquad \mathbf{F_b} = \begin{bmatrix} \mathbf{F_u} \\ \cdots \\ \mathbf{F_J} \end{bmatrix}.$$

An N by m matrix of actual criterion scores is denoted as $\mathbf{Z_u}$ and an $N$ by $n$ matrix of predictor scores in the joint predictor-criterion space is called $\mathbf{Y_J}$.

$$\mathbf{X_m}\ (\mathbf{F_u}' \mid \mathbf{F_J}') = (\mathbf{Z_u} \mid \mathbf{Y_J});\ (\mathbf{Y_J} \mid \mathbf{Z_u})'(\mathbf{Z_u} \mid \mathbf{Y_J}) = \mathbf{F_b}\mathbf{F_b}'.$$

Using the same $\mathbf{X_m}$ matrix as above and the transformation matrices described below, the relationships across the variables in $\mathbf{Z_u}$, $\mathbf{Z}$, and $\mathbf{Y_J}$ are preserved, as if the scores for all of these variables were obtained on the same sample of individuals. The $\mathbf{Z_u}$ matrix should be used only for evaluation of simulation results, while simulation decisions should be made on the basis of $\mathbf{Z}$ and/or $\mathbf{Y_J}$. Formulae for generating these three sets of variables are as follows:

$$\mathbf{Z_u} = \mathbf{X_m}\ \mathbf{R_u}^{1/2} \tag{13}$$

$$\mathbf{Z} = \mathbf{X_m}\ (\mathbf{V}\ \mathbf{R}_t^{-1/2})' \tag{14}$$

$$\mathbf{Y_J} = \mathbf{X_m}\ (\mathbf{F_{tp}})' \quad . \tag{15}$$

## APPENDIX 4B.5: GENERATING SETS OF PARALLEL FORMS

It is shown in Appendix 4C that the ability to generate two sets of parallel predictors with prescribed correlation coeficients between each pair of parallel forms can be the first step in the determination of the unbiased validity of a complexly determined predictor composite. It is difficult to otherwise obtain an unbiased estimate of the validity, and the standard error of estimate of this estimate, in cross samples. Obtaining these estimates is facilitated by (and/or probably requires) a model sampling experiment involving parallel forms when the development of predictor composites has included such procedures as selection of predictors for inclusion and the best weighting of these selected variables.

The prescribed correlation coefficients between parallel forms are, of course, one type of reliability coefficient. Although test developers frequently attend only to the relationships between pairs of parallel forms, when there are two sets of parallel forms the matrix of correlations between the two sets of parallel forms is relevant to their expected behavior in cross samples. Thus, we define a matrix $\mathbf{R_L}$ whose diagonal elements are the reliability coefficients, $r_{ii}$, and has the relationship to $\mathbf{R}_t$ indicated in the following equation:

$$\mathbf{R}_{tt} = \left[\begin{array}{c:c} \mathbf{R}_t & \mathbf{R_L} \\ \hdashline \mathbf{R_L} & \mathbf{R}_t \end{array}\right] .$$

where $\mathbf{R}_t$ and $\mathbf{R}_L$ are both $n$ by $n$ matrices. Different approaches to the creation of two sets of parallel forms will result in different values for the off diagonal elements of $\mathbf{R}_L$, although the diagonal elements always consist of the reliability coefficients ($r_{ii}$). We describe three different approaches for generating parallel forms such that,

$$[(\mathbf{Y}_{La} \mid \mathbf{Y}_{Lb})' (\mathbf{Y}_{La} \mid \mathbf{Y}_{Lb})/N] = \mathbf{R}_t.$$

Each approach implies a different concept of what constitutes sets of parallel forms.

We first describe a method of generating parallel forms which assumes that the elements of $\mathbf{R}_t$ are the correlations among true scores. Under this assumption the matrix $\mathbf{R}_L$ is equal to $\mathbf{S}_r \mathbf{R}_t \mathbf{S}_r$, where $\mathbf{S}_r$ is a diagonal matrix whose diagonal elements are the square roots of $r_{ii}$. That is, the off diagonal elements of $\mathbf{R}_L$ are equal to $(r_{ii})^{1/2}(r_{ij})(r_{jj})^{1/2}$ and the diagonal elements are equal to $r_{ii}$.

Denoting the factor solution of $\mathbf{R}_{tt}$ as the $2n$ by $k$ matrix $\mathbf{F}_{tt}$, we can say that by definition, $\mathbf{F}_{tt} \mathbf{F}_{tt}' = \mathbf{R}_{tt}$. Thus the $N$ by $2n$ matrix of parallel predictor scores, $(\mathbf{Y}_{La} \mid \mathbf{Y}_{Lb})$, can be generated from the relationship , $\mathbf{X}_{2n} (\mathbf{F}_{tt})' = (\mathbf{Y}_{La} \mid \mathbf{Y}_{Lb})$. These parallel sets of predictor scores are based on a traditional model of reliability and parallel forms which assumes that the Spearman-Brown formulae applies to the relationships of correlations between pairs of variables and their respective reliabilities.

Another model of parallel forms can be described by defining $\mathbf{F}_{tt}$ (the correlation between corresponding measures in the two sets of parallel forms still equal to $r_{ii}$) is as follows:

$$\mathbf{F}_{tt} = \left[\begin{array}{c:c:c} \mathbf{F}_L & \mathbf{O} & \mathbf{S}_L \\ \hdashline \mathbf{F}_L & \mathbf{S}_L & \mathbf{O} \end{array}\right] ,$$

where **0 is an *n* by 1 matrix of zeros,** $\mathbf{S}_L$ is a diagonal matrix whose diagonal elements are equal to $(1 - r_{ii})^{1/2}$, and $\mathbf{F}_L$ is as described below.

$\mathbf{R}_L$ is equal to $R_t - (S_L)^2$, and $\mathbf{F}_L \mathbf{F}_L' = \mathbf{R}_L$. We generate $\mathbf{Y}_{La}$ and $\mathbf{Y}_{Lb}$ using three separate $N$ by $n$ matrices of random normal deviates: $\mathbf{X}_L$, $\mathbf{X}_{La}$, and $\mathbf{X}_{Lb}$. We can generate these two sets of parallel forms as shown below:

$$\mathbf{Y}_{La} = (\mathbf{X}_L \mid \mathbf{X}_{La}) (\mathbf{F}_L \mid \mathbf{S}_L)' \tag{16a}$$

$$\mathbf{Y}_{La} = (\mathbf{X}_L \mid \mathbf{X}_{Lb}) (\mathbf{F}_L \mid \mathbf{S}_L)' \ . \tag{16b}$$

It is readily seen that in this approach $\mathbf{R}_L$ is equal to $\mathbf{R}_t$ with reliabilities replacing ones in the diagonals, and thus $\mathbf{F}_L$ is the traditional $n$ by $k$ initial factor solution in reliable space as contrasted with the solution in total test space utilized in the previous approach. The difference between ($\mathbf{F}_L \mathbf{F}_L'$) and $\mathbf{R}_t$ is assumed to be due to the error variance in the latter. The off-diagonal elements of the intercorrelations among predictors within the same set of parallel forms are accordingly equal to those found in the intercorrleations among tests across sets of parallel forms. This is the approach to be used when parallel sets of factor scores based on factors defined in the reliable space are to be generated, or if all the parellel forms are independently developed and only later randomly assigned to a particular set.

As noted above, test developers frequently pay attention only to the values in the diagonals of $\mathbf{R}_L$, in addition to the values of $\mathbf{R}_t$ and $\mathbf{V}$, in making a determination of whether two sets of predictors are sufficiently parallel. Sets of parallel forms developed under widely different conditions with respect to time, criteria, sample characteristics, etc., can have significant sources of error that creates an error variable correlated within the set, although uncorrelated with all other variables external to the set. Parallel forms can be generated which have this pattern of correlated error by defining $\mathbf{R}_L$ as follows:

$$\mathbf{F}_{tt} = \left[\begin{array}{c:c:c} \mathbf{F}_L & \mathbf{L} & \mathbf{O} \\ \hdashline \mathbf{F}_L & \mathbf{O} & \mathbf{L} \end{array}\right],$$

where $\mathbf{L}$ is an $n$ by 1 matrix whose elements are equal to the corresponding diagonal elements of $\mathbf{S}_L$, and $\mathbf{R}_L$ is defined as ($\mathbf{R}_t - \mathbf{LL'}$). The same $N$ by $n$+2 matrix of random normal deviates, $\mathbf{X}_{n+2}$, can be used to generate $\mathbf{Y}_{La}$ and $\mathbf{Y}_{Lb}$ as shown below:

$$\mathbf{X}_{n+2}\,(\mathbf{F}_{La})' = \mathbf{Y}_{La}\,;\ \mathbf{F}_{La} = (\mathbf{F}_L \mid \mathbf{L} \mid \mathbf{0}); \tag{17a}$$

$$\mathbf{X}_{n+2}\,(\mathbf{F}_{Lb})' = \mathbf{Y}_{Lb}\,;\ \mathbf{F}_{La} = (\mathbf{F}_L \mid \mathbf{O} \mid \mathbf{L}); \tag{17b}$$

$$E[(\mathbf{Y}_{La}'\mathbf{Y}_{La})/N] = E[(\mathbf{Y}_{Lb}'\mathbf{Y}_{Lb})/N] = \mathbf{R}_t;$$

$$E[(\mathbf{Y}_{La}'\mathbf{Y}_{Lb})/N] = \mathbf{R}_L\ .$$

# APPENDIX 4C
# RESEARCH DESIGN CONSIDERATIONS: SOME APPROACHES ESPECIALLY RELEVANT TO MODEL SAMPLING EXPERIMENTATION

In this appendix we consider a number of research design issues of special importance to the application of model sampling experimentation to the study of the effects on MPP of specific personnel policies or of assignment-classification methodology. The concept of using independent analysis and validation samples is not sufficient to control such biasing factors as predictor selection and/or provide accurate estimates of the standard error of estimate of the MPP resulting from personnel system policies. We have the means, through the use of model sampling techniques of controlling sources of bias that are left untouched in traditional empirical studies involving limited numbers of applicants or job incumbents that can be tested and evaluated, and divided into independent "back" and "cross" samples.

The ease with which the generation of analysis, evaluation, and cross samples, as well as sets of parallel forms of predictors and perturbed sets of utility measures, can be accomplished provide feasibility to the use of new creative research designs. We propose using the ease with which so many completely independent and representative samples of synthetic entities can be readily generated, to both control and measure effects of each type of bias.

We consider several types of bias we believe to be of particular interest to investigators of utility in personnel selection and classification. We first examine the following sources of bias:

(a) Cross sample shrinkage of selected and optimally weighted predictor composites--this bias source includes selecting of tests for inclusion in a composite prior to the "best" weighting of these tests to form a test composite.

(b) The presence of correlated error due to optimal weights obtained from the same "back" sample being used to define: (1) selection, assignment or other personnel system decision variables serving as independent variables in a model sampling experiment--as well as (2) being used to define the evaluation variable from which MPP is computed.

(c) The use of job "value" weights derived for computing performance benefits (for inclusion in a utility estimate) in the determination of scores for selection, assignment, and other decision variables; there is obviously something suspect in the inclusion of the dependent variable as one of the decision variables.

The control of the first of these bias sources can significantly increase the utility of operational procedures. The latter two have their primary importance in their potential for the erroneous inflation of utility estimates.

The first of our three bias sources is evidenced by the traditional shrinkage of multiple correlation coefficients when "best" weights are applied to scores in samples independent of the analysis samples on which the "best" weights were computed. The formulae for obtaining an estimate of the expected, unbiased validity of the "best" weighted composite in an independent sample will, of course, be applicable only when certain assumptions are met. The best known one, as proposed by Wherry (Catlin,1980), assumes that the "best" weights were computed in the universe, rather than in another independent sample drawn from that universe. The results of such a formulae are appropriately compared with a model sampling experiment in which the regression weights are computed using the values of $\mathbf{R}_t$ and $\mathbf{V}$ designated by the investigator as defining the universe. These universe weights are applied to scores of entities generated from the designated universe for as many cross samples as desired, and good estimates of both the expected, unbiased validity and the standard deviation of these estimates across samples are available.

None of the shrinkage formulae provide good estimates when test selection has preceded the computation of regression weights. In a model sampling experiment to determine shrinkage under conditions that include test selection, the test selection procedures and computation of weights should be computed using successive independent samples of entities.

Methods for adjusting regression weights to compensate for the effect of sampling error has the effect of reducing the range (i.e., variance) of the weights across predictors. One such method is called ridge analysis (Draper and Van Nostrand, 1979). The benefits of using such adjustments cannot be evaluated by comparing in each cross sample the validities of composites defined using different methods of adjusting weights computed from a number of different analysis samples. A model sampling experiment cannot help in the obtaining of better estimates of the universe parameters.

It might appear that a reasonable goal for adjusting regression weights would be to obtain both the largest mean validity and the tightest cluster of estimates around the mean validity. Unfortunately neither of these goals is obtainable from regression weight adjustment methods that rely on reducing the differences (causing a flattening effect) among regression weights within a composite. The formulae for adjusting regression weights by making them more similar across the independent variables provides a means of changing the estimated universe values, not a means of reducing the shrinkage of validities in cross samples.

While an investigator could use model sampling to produce similar adjustments to regression weights as would be produced by these formulae, we would not expect the model sampling results in independent samples to show a reduction in shrinkage. The lower validity obtained in cross samples would be the result of lowering the magnitude of the "back" validity, usually by systematically increasing the effect of the general factor at the expense of the group factors. Thus, the use of these formulae in personnel research falls under the heading of how to best estimate universe parameters rather than how to correct for shrinkage effects.

An adjustment (we hesitate to call it a correction) for inflated differences among regression weights computed using empirically obtained estimates of universe values for $\mathbf{R}_t$ and $\mathbf{V}$ is obtainable by providing parameters for producing factor scores from one independent sample ($\mathbf{Y}(\mathbf{R}_t)^{-1}\mathbf{F}_{tt} = \mathbf{Q}_t$) and then computing the regression weights for predicting performance from these factor scores ($\mathbf{Q}_t\mathbf{W}_t' = \mathbf{Z}$) on a second independent analysis sample. The conversion of regression weights for factor scores to regression weights for personnel tests would in itself have the effect of flattening, reducing the variance of, the regression weights, because of the reduction in the dimensionality of the joint predictor-criterion space.

The effect of computing weights on one set of parallel forms to be applied operationally on the other set in independent samples can also be readily evaluated in a model sampling experiment (see Appendix 4B.3 on the generation of sets of parallel forms).

We have repeatedly emphasized the importance of producing as the final output of a model sampling experiment an estimate of MPP based on all information available to the experimenter. This MPP is the average predicted performance (PP) resulting from the decisions made in a simulation of portions of a personnel system.

The sample on which the parameter values required to compute PP variables for evaluation purposes, that is to compute the final MPP standard score, is called the evaluation sample, although the actual evaluative results are obtained from applying these parameters to synthetic scores generated for inclusion in a "cross" sample.

In some model sampling experiments the investigator can assume that the universe parameters are completely known and the experiment is thus a means of solving a mathematical problem that currently eludes solution by analytical means. However, when research objectives require consideration of inflated relationships due to test selection, best weighting of predictors, or other decisions that capitalize on sampling error, the elimination of correlated error present in both decision variables and evaluation variables becomes an important aspect of the experiment. Since such correlated error will result when the parameters defining decision variables are computed on the same analysis sample as the parameters defining the evaluation variables, this type of correlated error is eliminated by assuring that selection, weighting, etc., for decision and evaluation variables are accomplished on independent samples.

In most model sampling experiments relating to the utility of selection and classification of personnel it is very important that the same estimate of PP be utilized as the evaluation variable, regardless of the experimental conditions or the different variables (the selection and assignment variables, etc.) that may be utilized in the implementation of the various experimental conditions. When one of the decision processes being simulated is optimal assignment of personnel to jobs, it is generally inappropriate to use the objective function (i.e., the variable being maximized, frequently called the allocation sum in dual LP programs for assignment of personnel) as the measure of MPP. Instead, the overall "best" estimate of PP should be used to compute MPP at the end of the simulation of the personnel decision processes--independently of the variables used in the selection, assignment, and other decision processes of the simulation.

The steps of a model sampling experiment can be divided into three parts that will often require independent samples for their implementation. The estimates of $\mathbf{R}_t$ and $\mathbf{V}$ designated as universe values should be used to compute the post multipliers of $\mathbf{X}$ to generate the synthetic scores for predictor, criterion, and other decision variables for use in the cross samples. One or more independent "analysis" samples are used to compute the parameters that are applied to cross sample scores to produce unbiased estimates of predictor composites (selection and assignment variables) and other system decision variables (e.g., for determining promotion, length of training, school grades, attrition, etc.)

In methodological model sampling studies where alternative methodologies or policies are being compared, it makes sense to to use the "universe" sample as the source of evaluation parameter values. On the other hand, when the objective af a model sampling experiment is to provide a statistical basis for interpreting the results of an empirical research study, the investigator will wish to closely emulate the division of available empirical data into independent samples, drawing separate analysis, evaluation and cross samples of entities of the same size and structure as in the empirical study; additional cross samples or larger cross samples are of course available to provide greater precision and can be used without disturbing the closeness with which the empirical study is replicated.

The most perplexing research design issue relating to the conduct of studies to estimate utility, for either analytical, empirical or potential model sampling experiments, arises when separate values are placed on performance in each job. In such studies, the evaluation variable has usually been a function of job predictability (validity), performance value for each job, and some measure of the performance spread of incumbents (possibly, the standard deviation of job performance in dollar terms, or $SD_y$). In future model sampling experiments the evaluation variable would most likely be the MPP for each job multiplied by the dollar value of performance at that level.

Selection and assignment decisions are often made on the basis of the predictability and value of jobs (sometimes directly and sometimes by making the standard deviation of predictors proportional to their validity), as well as the magnitude of each individuals test composite scores. Very often, a high percentage of the utility accruing to a personnel system component is due to the presence of the same job value scores as components in both the decision and evaluation variables.

Job value scores are commonly based on the judgments of middle to high level managers. There is often a tendency to believe the validity of such judgments to be perfectly correlated with the managerial level of those making these judgments. This belief may make the concept of inter-rater reliability difficult to put in practice. Also, the process by which these judgments are obtained have the potential for adding considerable method bias to the scores, and high level managers are generally not available for research on the effects of alternative methods. In summary, selection-assignment systems using job values on the system decision side should not be justified by utility studies that use the same job value scores on the evaluation side. Appropriate utility studies of such systems may have to await the time when much more is known about obtaining two truly independent estimates of job value scores, one for use on the decision side and the other on the evaluation side.

# GLOSSARY

**ability test**[a]--A test that measures the current performance or estimates future performance of a person in some defined domain of cognitive, psychomotor, or physical functioning.

**achievement test**[a]--A test that measures the extent to which a person commands a certain body of information or possesses a certain skill, usually in a field where training or instruction has been received.

**adaptive testing**[a]--A sequential form of testing in which successive items in the test are chosen based on the responses to previous items.

**algebraic variability derivation**--A technique for incorporating uncertainty into utility by the use of variance estimates.

**allocation efficiency**--The gain in benefit over random assignment obtained from an optimal assignment process attributable to differential validity.

**allocation process**--Classification that capitalizes on differential job validity.

**alternative**[c]--A course of action whose selection may result in an outcome that will attain the original objective.

**aptitude test**[a]--A test that estimates future performance on other tasks not necessarily having evident similarity to the test tasks. Aptitude tests are often aimed at indicating an individual's readiness to learn or to develop proficiency in some particular area if education or training is provided. Aptitude tests sometimes do not differ in form or substance from achievement tests, but may differ in use and interpretation.

**assessment procedure**[a]--Any method used to measure characteristics of people, programs, or objects.

**attenuation**[a]--The reduction of a correlation or regression coefficient from its theoretical true value due to the imperfect reliability of one or both measures entering into the relationship.

**battery**[a]--A set of tests standardized on the same population, so that norm-referenced scores on the several tests can be compared or used in combination for decision-making.

**behavior**[b]--Observable aspects of a person's activities.

**benefit**--A theoretically desirable measure of performance that is value-weighted for jobs and validity in terms of an appropriate metric; when the benefit measure is correctly combined with costs, it provides a measure of utility.

**break-even values**--The determination of the lowest value of any individual parameter that would still yield a positive total utility value.

**classification**--The matching of individuals and jobs in an organization with the goal of maximizing aggregate performance; it requires multiple predictors jointly measuring more than one dimension and multidimensional job criteria.

**classification**[a]--The act of determining which of several possible job assignments a person is to receive.

**classification battery**-- A battery of tests used operationally to classify personnel.

**classification efficiency**--The gain in benefits over random assignment obtained from an optimal assignment process attributable to allocation and hierarchical classification efficiency; a separate LSE must be used for each criterion.

**cognition**[c]--The act or process of knowing, including both awareness and judgment.

**composite score**[a]--A score that combines several scores by a specified formula.

**concurrent criterion-related validity**[a]--Evidence of criterion-related validity in which predictor and criterion information are obtained at approximately the same time.

**construct**[a]--A psychological characteristic (e.g., numerical ability, spatial ability, introversion, anxiety) considered to vary or differ across individuals. A construct (sometimes called a latent variable) is not directly observable; rather it is a theoretical concept derived from research and other experience that has been constructed to explain observable behavior patterns. When test scores are interpreted by using a construct, the scores are placed in a conceptual framework.

**cost accounting approach**--The approach used to develop a dollar criterion that considers the value of products and services and the organization's costs to provide products and services.

**cost effectiveness**[c]--A state or condition in which the benefits associated with a particular outcome clearly exceed the cost of obtaining the outcome.

**decision**[c]--A moment of choice in an ongoing process of evaluating alternatives with a view to selecting one or some combination of them to attain the desired end.

**decision tree**[c]--A framework for developing the anatomy of a decisionmaking situation that uses the concepts of probability, utility, and expected value.

**decision theoretic approach**--The set of alternatives, costs and possible outcomes leading to a choice.

**differential validity**--The level of prediction using LSEs of differences among criterion scores when referring to $H_d$; this measure is related to the variation of a validity vector with jobs and to an assignment variable being more valid for its own job family than any other job family.

**discounting**--A procedure for equating the costs and benefits that accrue over time to reflect the opportunity costs and returns foregone.

**efficiency**--A solution that minimizes costs as measured by physical resources and time utilized.

**expected value**[c]--A concept that permits a decisionmaker to place a monetary or other value on the positive and negative consequences likely to result from the selection of a particular alternative.

**external employee movement**--The analysis of employee separations and acquisitions in an organization.

**goal**[c]--A subset of an objective expressed in terms of one or more specific dimensions.

**gross national product**--The sum of all expenditures on goods and services by households, by firms on new capital, and by government.

**hierarchical classification efficiency**--All classification efficiency not explainable as allocation efficiency; it capitalizes on disparate variances of the mean predicted benefit scores for the corresponding jobs.

**hierarchical layering**--A phenomenon in which LSEs are more valid or of more value for some jobs than for others.

**human capital**--The skills of the workforce that determine what workers can contribute to the production process.

**human resource accounting**--The economic consequences of employees' behavior.

**inter-rater reliability**[a]--Consistency of judgments made about people or objects among raters or sets of raters.

**interest inventory**[a]--A set of questions or statements that is used to infer the interests, preferences, likes, and dislikes of a respondent.

**inventory**[a]--A questionnaire or checklist, usually in the form of a self-report, that elicits information about an individual. Inventories are not tests in the strict sense; they are most often concerned with personality characteristics, interests, attitudes, preferences, personal problems, motivation, and so forth.

**item analysis**[a]--The process of assessing certain characteristics of test items, usually the difficulty value, the discriminating power, and sometimes the correlation with an external criterion.

**job analysis**[a]--Any of several methods of identifying the tasks performed on a job or the knowledge, skills, and abilities required to perform that job.

**job relatedness**[b]--The inference that scores on a selection instrument are relevant to performance or other behavior on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the relevance of the content of the selection instrument, or of the construct measured.

**joint probability**[c]--The probability that two or more events will occur.

**labor**--The worker effort available to the production process.

**law of diminishing returns**--As the quantity of an input is increased and the quantity of other inputs stays the same, a point is reached where the additional output produced per unit of added input declines.

**linear combination**[b]--The sum of scores, whether weighted differentially or not, on different assessments to form a single composite score.

**linear model**[c]--A model of choice in which the evaluation of each alternative is based on the sum of its weighted values on all its dimensions, and the alternative with the greatest sum is the obvious choice.

**longitudinal study**[a]--Research that involves the measurement of a single sample at several different points in time.

**marginal cost**--The cost of producing an additional unit.

**maximizing behavior**[c]--An approach to decisionmaking oriented toward obtaining an outcome of the highest quantity or value.

**mean predicted performance (MPP)**--The measurement of benefits can be approximated by computing MPP across jobs; if MPP is weighted by the value of each job, it becomes a more useful measure of benefits. It provides a means of comparing the effectiveness of alternative tests or test batteries in the context of a specified set of jobs and performance scores.

**meta-analysis**[b]--A procedure to cumulate findings from a number of validity studies to estimate the validity of the procedure for the kinds of jobs or groups of jobs and settings included in the studies.

**meta-analysis**--A technique for determining the degree to which the variance in validity coefficients across situations for job-test combinations is due to statistical artifacts.

**model**[c]--A physical or abstract representation of some part of the real world that is used to describe, explain, or predict behavior.

**Monte Carlo analysis**--A stochastic technique that can provide numerical solutions for mathematical functions lacking analytic solutions; the analysis typically uses random numbers as input to an evaluation process employing variance reduction procedures.

**multidimensional screening (MDS)**--A selection/classification process using an algorithm that ensures no nonselected person has a higher predicted performance on any job than the person assigned to that job; the algorithm also ensures that no other assignment can further raise the mean predicted performance.

**multivariate**[b]--Characterizing a measure or study that incorporates several variables.

**norms**[a]--Statistics or tabular data that summarize the test performance of specified groups, such as test takers of various ages or grades. Norms are often assumed to represent some larger population, such as test takers throughout the country.

**norm-referenced test**[a]--An instrument for which interpretation is based on the comparison of a test taker's performance to the performance of other people in a specified group.

**objective**[b]--Pertaining to scores obtained in a way that minimizes bias or error due to different observers or scores.

**operational efficiency**--The improvement in MPP obtained from the usually imperfect operational selection assignment process as contrasted to potential efficiency, the improvement obtainable if the maximally efficient prediction composites of a given battery were to be used in optimal selection/assignment algorithms.

**opportunity cost**[c]--The cost of the next best alternative that is sacrificed to select what appears to be the best alternative.

**payoff**[c]--The intersection of an alternative and a state of nature in a payoff table; it measures the value (utility) to the decisionmaker likely to result from the selection of that alternative given the probabilistic occurrence of the state of nature.

**payoff table**[c]--A convenient framework in which to present the elements of a decision making situation employing the concepts of probability, utility, and expected value.

**percentile**[a]--The score on a test below which a given percentage of scores fall.

**performance**[b]--The effectiveness and value of work behavior and its outcomes.

**personality inventory**[a]--An inventory that measures one or more characteristics that are regarded generally as psychological attributes or interpersonal skills.

**placement**--A procedure in which individuals are matched to levels within jobs as contrasted to the classification process of matching personnel to jobs.

**potential allocation efficiency**--The maximum allocation effectiveness achievable from the differential validity of a given test battery and set of jobs expressed as a mean predicted performance standard score.

**potential classification efficiency**--The maximum classification effectiveness achievable from a given test battery and set of jobs expressed as a mean predicted performance standard score; it incorporates both potential allocation efficiency and hierarchical layering effects.

**potential selection efficiency**--Rank-ordering applicants on some benefit continuum and rejecting all those below some point on that continuum.

**potential utilization efficiency**--The sum of potential selection efficiency and potential classification efficiency.

**predictive criterion-related validity**[a]--Evidence of criterion-related validity in which criterion scores are observed at a later date, for example, for job or school performance.

**predictor**[a]--A measurable characteristic that predicts criterion performance such as scores on a test, evidence of previous performance, and judgments of interviewers, panels, or raters.

**productivity**--The ratio of outputs to inputs of a resource (workers, capital equipment); a measure of the degree of the use of resources.

**psychometric**[a]--Pertaining to the measurement of psychological characteristics such as abilities, aptitudes, achievement, personality, traits, skill, and knowledge.

**regression equation**[b]--An algebraic equation used to predict criterion performance from predictor scores.

**relevance**[b]--The extent to which a criterion measure reflects important job performance dimensions or behaviors.

**reliability**[a]--The degree to which test scores are consistent, dependable, or repeatable, that is, the degree to which they are free of errors of measurement.

**reliability coefficient**[a]--The square of the correlation of an observed score with its "true" component; often measured as the coefficient of correlation between two administrations of a test. The conditions of administration may involve variation of test forms, raters or scorers, or passage of time. These and other changes in conditions give rise to qualifying adjectives being used to describe the particular coefficient, e.g., parallel form reliability, rater reliability, test retest reliability, etc.

**residual score**[a]--The difference between the observed and the true or predicted score.

**restriction of range**[a]--A situation in which, because of sampling restrictions, the variability of data in the sample is less than the variability in the population of interest.

**risk**[c]--A common state or condition in decision making characterized by the possession of incomplete information regarding a probabilistic outcome.

**sample**[b]--The individuals who are actually tested from among those in the population to which the procedure is to be applied.

**score**[a]--Any specific number resulting from the assessment of an individual; a generic term applied for convenience to such diverse measures as test scores, estimates of latent variables, production counts, absence records, course grades, ratings, and so forth.

**selection**--A procedure for rejecting some applicants for organizational membership as contrasted to assigning all applicants to jobs (classification); or rejecting an applicant for a single job as contrasted to selection and assignment to one of a number of jobs (multidimensional selection).

**selection decision**[a]--A decision to accept or reject applicants for a job on the basis of information.

**selection instrument**[b]--Any method or device used to evaluate characteristics of persons as a basis for accepting or rejecting applicants.

**selection procedures**[b]--Process of arriving at a selection decision.

**sensitivity analysis**--An analytic technique in which a utility parameter is varied through a range of values, holding other parameter values constant to determine the impact on the total utility estimates.

**shrinkage**[a]--Refers to the fact that a prediction equation based on a first sample will tend not to fit a second so well.

**shrinkage correction**[b]--Adjustment to the multiple correlation coefficient for the fact that the beta weights in a prediction equation cannot be expected to fit a second sample as well as the original.

**simulation model**[c]--A special type of abstract model that is analogous to a segment of the real world and contains a time dimension. It is used to explain and predict behavior as if it occurred in the real world.

**skill**[b]--Competence to perform the work required by the job.

**split-half reliability coefficient**[a]--An internal analysis coefficient obtained by using half the items on the test to yield one score and the other half of the items to yield a second, independent score. The correlation between the scores on these two half-tests, stepped up via the Spearman-Brown Formula, provides an estimate of the alternate-form reliability of the total test.

**standard score**[a]--A score that describes the location of a person's score within a set of scores in terms of its distance from the mean in standard deviation units.

**standardized prediction**[b]--A test employed for estimating a criterion of job performance, the test having been developed and normative information produced according to professionally prescribed methods as described in standard reference works.

**standards**[c]--Criteria against which the results of an implemented decision can be measured.

**state of nature**[c]--A state or condition likely to prevail when a choice is made.

**sunk costs**--Costs that once incurred cannot be changed by future action.

**test**[b]--A measure based on a sample of behavior.

**test fairness**--The most commonly accepted model of test fairness is the regression model; a fair test predicts the job performance of a minority and the majority in the same way.

**test-retest coefficient**[a]--A reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores.

**trade-off value**[c]--A value that exists when a given amount of one kind of performance may in some measure be substituted for another kind of performance.

**traditional selection approach**--The view of tests as measuring instruments intended to assign accurate values to attributes of an individual stressing precision of measurement and estimation rather than selection outcomes.

**unidimensionality**[a]--A characteristic of a test that measures only one latent variable.

**utility**[c]--Technically, want-satisfying power; it is often defined as the preference of the decisionmaker for a given outcome.

**utility analysis**--The determination of institutional gain or loss (outcomes) anticipated from various courses of action usually measured in terms of dollars.

**validity**[a]--The degree to which a certain inference from a test is appropriate or meaningful.

**validity coefficient**[a]--A coefficient of correlation that shows the strength of the relation between predictor and criterion.

**validity generalization**[a]--Applying validity evidence obtained in one or more situations to other similar situations on the basis of simultaneous estimation, meta-analysis, or synthetic validation arguments.

**values**[c]--The nominative standards by which human beings and organizations are influenced in their choices.

**variability**[b]--The spread or scatter of scores.

**variable**[a]--A quantity that may take on any one of a specified set of values.

**variance**[a]--A measure of variability; the average squared deviation from the mean: the square of the standard deviation; and, in the experimental design literature, the sum of the squared deviation from its mean doubled by the degrees of freedom.

**Z-score**[a]--A type of standard score scale in which the mean equals zero and the standard deviation equals one unit for the group used in defining the scale.

## NOTES:

a Adapted from American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). *Standards for Education and Psychological Testing.*

b Adapted from Society for Industrial and Organization Psychology (1987). *Principles for the Validation and Use of Personnel Selection Procedures.*

c Adapted from Heyne (1988). *Microeconomics.*

# REFERENCES


Abbe, E.N. (1968, December). *Statistical properties of allocation averages.* Research Memorandum 68-13. Statistical Research and Analysis Division. Washington, DC: U.S. Army Behavioral Science Research Laboratory.

Alf, E.F. and Wolfe, (1968, June). *Comparison of classification strategies by computer simulation methods.* Technical Bulletin STB 68-11. San Diego, CA: U.S. Naval Personnel Research Activity.

Anastasi, A. (1982). *Psychological Testing* (5th Ed.). New York: Macmillan.

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* New York: Academic Press.

Barr, Donald R. and Slezak, Norman L. (1972, December). A comparison of multivariate normal generators. *Communication of American Computing Machinery*, **12**, 1048-1049.

Boldt, R.F. and Johnson, C.D. (1963). *Computerized personnel assignment system.* Paper presented at the Second Army Operations Research Symposium. Dunn, NC.

Brogden, H.E. (1946a). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, **37**, 65-76.

Brogden, H.E. (1946b). An approach to the problem of differential prediction. *Psychometrika*, **11**, 139-154.

Brogden, H.E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement*, **11**, 173-196.

Brogden, H.E. (1954a). A simple proof of a personnel classification theorem. *Psychometrika*, **19**, 205-208.

Brogden, H.E. (1954b). Statistical theory and research design. *Annual Review of Psychology*, **V**, 377-400.

Brogden, H.E. (1955). Least squares estimates and optimal classification. *Psychometrika*, **20**, 249-252.

Brogden, H.E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement*, **19**, 181-190.

Brogden, H.E. (1964). Simplified regression patterns for classification. *Psychometrika*, **29**, 393-396.

Canavos, G.C. (1967). A comparative analysis of two concepts in the generation of uniform pseudo-random numbers. *Proceedings of the 1967 National Meeting of the A.C.M.*, 485-501.

Cardinet, J. (1959). The use of profiles for differential classification. *Education and Psychological Measurement*, **XIX**, 191-205.

Cascio, W.F. (1987b). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Catlin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology*. **65**. 407-414.

Cronbach, L.J., and Gleser, G.C. (1965). *Psychological tests and personnel decisions*. Second Edition. Urbana: University of Illinois Press. Originally published 1957.

Draper, N.R. and Van Nostrand, R.C. (1979). Ridge regression and James-Stein estimation: Review and comments. *Technometrics*, **21** (4), 451-466.

Dwyer, P.S. (1937). The determination of factor loadings of a given test from the known factor loadings of other tests. *Psychometrika*, **11**, 173-178.

Dwyer, P.S. (1954). Solution of the personnel classification problem with the method of optimal regions. *Psychometrika*, **18**, 11-26.

Dwyer, P.S. (1957). The detailed method of optimal regions. *Psychometrika*, **22**, 43-52.

Fishman, G.S. and Moore, L.R. (1986). An exhaustive analysis of multiplicative congruential random number generators with modulus $2^{31}-1$. *SIAM Journal of Science and Statistical Computing*, **7**(1), 24-45.

Granda, T.M. and McMullen, R.L. (1974). *A heuristic approach to a large personnel assignment problem*. Paper presented at the Fall ORSA/TIMS National Convention.

Green, B.F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, **17**, 433-440.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Harris, R.N. (1966). *Multiple significance tests of correlation coefficients in correlation matrices*. Dissertation, Order No. 67-11,302. College Park: University of Maryland.

Harris, R.N. (1967, March). *A model sampling experiment to evaluate two methods of test selection*. (Research Memorandum, 67-2), Statistical Research and Analysis Division. Washington, DC: U.S. Army Behavior and Systems Research Laboratory.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: John Wiley.

Hatch, R.S. (1966). *ACT II System: System Memorandum*. Rockville, MD: Decision Systems Associates, Inc.

Hatch, R.S. (1970). Development of optimal allocation algorithms for personnel assignment. In A.R. Smith (Ed.), *Models of Manpower Systems*. London: English University Press.

Herzberg, P.A. (1969). The parameters of cross-validiation. *Psychometrika*, **34** (Monograph Supplement, Monograph No. 16), 1-70.

Horst, P. (1954). A technique for the development of a differential prediction battery. *Psychological Monographs*, **68**(9, Whole No. 380).

Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. *Psychological Monographs*, **69**(5, Whole No. 380), 1-22.

Horst, P. (1956a). Multiple classification by the method of least squares. *Journal of Clinical Psychology*, **12**, 3-16.

Horst, P. (1956b). Optimal test length for maximum differential predictions. *Psychometrika*, **21**, 51-66.

Horst, P. (1960). Optimal estimates of multiple criteria with restrictions on the covariance matrix of estimated criteria. *Psychological Reports*, Monograph Supplement 6-V6, 427-444.

Horst, P., and MacEwan, C. (1956). Optimum test length for maximum absolute prediction. *Psychometrika*, **21**, 111-124.

Horst, P., and MacEwan, C. (1960). Predictor eliminator techniques for determining multiple prediction batteries. *Psychological Reports*, **7**, 19-50.

Horst, P. (1966). *Psychological measurement and prediction*. Belmont, CA: Wadsworth.

Hunter, J.E. (1983). *Validity generalization for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery (GATB)*. USES Test Research Report No. 45. Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, **29**, 340-362.

Hunter, J.E., and Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E.A. Fleishman and M.D. Dunnette (Eds.), *Human performance and productivity: Vol. 1: Human capability assessment*. Hillsdale, NJ: Erlbaum.

Johnson, C.D. (1970, February). *Computer program for transforming a factor solution, F, into a solution which maximizes the sum of the absolute distances among points representing criteria or other reference variables*. Internal Memoranda. Arlington, VA: U.S. Army Behavioral Systems Research Laboratory.

Johnson, C.D., Klieger, W.A., and Frankfeldt, E. (1958, April). *Construction of a classification self-description blank designed to emphasize differences among occupational areas*. Research Memorandum 58-6. Washington, DC: Personnel Research Branch, The Adjutant General's Office, Dept. of the Army.

Johnson, C.D., and Kotula, L.J. (1958, August). *Validation of experimental self-description materials for general and differential classification.* Technical Research Note 95. Washington, DC: Personnel Research Branch, The Adjutant General's Office, Dept. of the Army.

Johnson, C.D., and Sorenson, R.C.. (1974). Model sampling experimentation for manpower planning. In D.J. Clough, C.G. Lewis, and A.L. Oliver (Eds.), *Manpower Planning Models*, pp. 37-52. London: English Universities Press.

Kaiser, H.F., and Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, **27**, 179-182.

Knuth, D.E. (1981). *The art of computer programming, Vol. II: Seminumerical algorithms*. Reading, MA: Addison Wesley.

Kruskal, J. (1977). The relationship between multidimensional scaling and clustering. In J. Van Ryzin (Ed.), *Classification and Clustering*, 17-44. New York: Academic Press.

Kuehl, F.W. (1969, September). *Evaluation of a multiplicative generator of pseudo-random numbers*. Technical Research Note 215. Statistical Research and Analysis Division. Washington, DC: U.S. Army Behavioral Science Research Laboratory.

Larkin, W.D. (1965, May). *Computer generation of random numbers*. Research Memorandum 65-3. Statistical Research and Analysis Laboratory. Washington, DC: U.S. Army Personnel Research Office.

Larkin, W.D. (1966, November). *Evaluation of a modified optimal regions algorithm for manpower systems experiments*. Research Memorandum 66-9. Statistical Research and Analysis Laboratory. Washington, DC: U.S. Army Personnel Research Office.

Lewis, T.G. (1975). *Distribution sampling for computer simulation*. Lexington, MA: Lexington Books, D.C. Heath.

Lord, F.M. (1952). Notes on a problem of multiple classification. *Psychometrika*, **17**(3), 297-304.

Maclaren, M.D., and Marsaglia, G. (1965, January). Uniform random number generators. *Journal of American Computing Machinery*, **1**, 83-89.

McLaughlin, D.H., Rossmeissl, P.G., Wise, L.L., Brant, D.A., and Wang, M.M. (1984, October). *Validation of current and alternative ASVAB area composites, based on training and SQT information of FY81 and FY82 enlisted accessions*. Technical Report 651. Alexandria, VA: U.S. Army Research Institute.

McLaughlin, D.H. (1985, August). *Measurement of test battery value for selection and classification*. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Naylor, T.H., Balintfy, J.L., Burdick, D.S., and Chu, K. (1966). *Computer simulation techniques*. New York: Wiley.

Niehl, E., and Sorenson, R.C. (1968, January). *Simpo-I entity model for determining the qualitative impact on personnel policies*. Washington, DC: Technical Research Note 193. Statistical Research and Analysis Division, U.S. Army Behavior and Systems Research Laboratory. (AD 831-268).

Olson, P.T., Sorenson, R.C., Haynam, K.W., Witt, J.M., and Abbe, E.N. (1968, March). *Summary of SIMPO-I model development*. Statistical Research and Analysis Division. Washington, DC: U.S. Army Behavioral Research Laboratory.

Olson, P.T. (1968, August). *Use of Army school samples in estimating ACB test validity*. Statistical Research and Analysis Division. Washington, DC: U.S. Army Behavioral Science Research Laboratory.

Park, S.K., and Miller, K.W. (1988). Random number generators: good ones are hard to find. *Communications of the ACM*, **31**, 1192-1201.

Schmidt, F.L., Hunter, J.E., and Dunn, W.L. (1987, November). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB)*. Contract No. Delivery Order 0053. San Diego, CA: U.S. Navy Personnel Research and Development Center.

Shields, J.L. (1978, July). *An empirical investigation of the effect of heteroscedasticity and heterogeneity of variance on the analysis of covariance and the Johnson-Neyman Technique*. Technical Paper 292. Alexandria, VA: U.S. Army Research Institute.

Sneath, P.H., and Sokal, R.R. (1973). *Numerical taxonomy*. San Francisco: Freeman.

Sokal, Robert R. (1977). Clustering and classification: background and current directions. In J. Van Ryzin (ed.), *Classification and Clustering*, 1-15. New York: Academic Press.

Sorenson, R.C. (1965a, November). *Optimal allocation of enlisted men--full regression equations vs. aptitude area scores*. Technical Research Note 163. Washington, DC: U.S. Army Personnel Research Office. (AD 625 224).

Sorenson, R.C. (1965b, August). *Development and evaluation of a matrix transformation useful in personnel classification*. An unnumbered report printed by University of Washington, Seattle, WA, p. 111.

Tausworthe, R.C. (1965). Random numbers generated by linear recurrence modula two. *Mathematics of Computation*, **19**, 201-209.

Tippett, L.C.H. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, **XVII**, 364-387.

Van Ryzin, J. (ed.) *Classification and Clustering*. New York: Academic Press.

Ward, J.H. (1958). The counseling assignment problem. *Psychometrika*, **23**, 55-65.

Wherry, R.J., Sr., Naylor, J.C., Wherry, R.J., Jr., and Fallis, R.F. (1965). Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika*, **30**, 303-313.

Young, R.D. (1969). Primal integer progrmming. In T.C. Hu (Ed.), *Integer Programming and Network Flow*, Chapter 11. Reading, MA: Addison-Wesley.

Zeidner, J., and Johnson, C.D. (1989a, July). *The utility of selection for military and civilian jobs*. IDA Paper P-2239. Alexandria, VA: Institute for Defense Analyses.

Zeidner, J., and Johnson, C.D. (1989b, September). *The economic benefits of predicting job performance*. IDA Paper P-2241. Alexandria, VA: Institute for Defense Analyses.